

Gene therapy deserves a fresh chance

Initial interest in gene therapy waned after the technology failed to live up to expectation. Progress made since has received little attention, but suggests that the pervading sense of disillusionment is misplaced.

In the early 1990s, when the first human trials got under way, it seemed to many that the era of gene therapy was at hand: the techniques of modern molecular biotechnology would make it possible to repair genetic defects by inserting healthy DNA directly into a patient's cells. The excitement was short-lived. Lasting effects proved difficult to obtain in early trials, and the community quickly grew sceptical. Then, in 2003, when it was announced that several gene-therapy patients in a Paris-based clinical trial had developed leukaemia, and that one of them had died, the mood became bleak. Subsequent reports of successful and effective gene-therapy trials have done little to lift the prevailing sense of doom. For most researchers, gene therapy now seems like a dead end.

But it doesn't have to be a dead end — not if scientists shift their perspective on the risks of gene therapy to be more in line with that of clinicians.

Scientists are trained to focus on understanding the systems that they study in great detail. And when they devise therapeutic interventions — for example, harnessing a viral shell to insert a therapeutic gene into a patient's DNA — they naturally want those systems to be engineered with equally great care, and for them to be as near to risk-free perfection as possible.

Clinicians, by contrast, care for real patients in real time, which makes treatment decisions a matter of pragmatism. How do the risks stack up against the benefits for each available alternative — given that the risks are never zero? Clinicians are certainly not cavalier about their patients' well-being, but they may well end up prescribing a therapy that has a poorly understood mechanism and potentially large side effects because it gives the patient the best odds of recovery or survival. If they — and patients — had shied away from such dangers in the past, life-saving interventions such as organ grafts and bone-marrow transplants might never have been developed.

From that perspective, the fact that, collectively, the Paris trial and others carried out since have produced positive results in some 20 patients out of a total of two dozen leukaemia cases at least as large as the handful of leukaemia cases. To clinicians, such results suggest a treatment that is risky, but potentially life-saving — a new option for people for whom there are no alternatives.

However, this was not the view that prevailed. When the viral delivery vehicle itself turned out to be responsible for the leukaemia cases in the Paris trial, scientists deemed the trial a failure. Bad press ensued, proposals for gene-therapy clinical trials came under increased regulatory scrutiny and standards for demonstrating safety were set higher than for other approaches. Unsurprisingly in such a climate, the biotechnology and pharmaceutical industries gradually dropped out of the gene-therapy pursuit. This corporate disinterest slowed clinical progress: academic centres are ill-equipped to make gene-therapy vectors of clinical grade and scale, and research funding is typically insufficient to support clinical trials. More insidiously, it has become harder to recruit young talent to a field that is perceived as falling short of its promises.

To reverse this trend, it is time for researchers and industry to refresh their perspective on gene therapy and to consider its successes with as much intensity as its setbacks. The focus on adverse events has had positive consequences: researchers dissected the exact molecular mechanisms that led to cancer, designed better vectors, devised animal models to test these vectors and developed sophisticated assays for monitoring patients. As a result, both scientists and clinicians now have a battery of extraordinarily refined tools for preclinical and clinical studies of gene therapy. The field is ripe for further successes. ■

"The results suggest a treatment that is risky, but potentially life-saving."

Darwin and culture

A new series of essays traces the astounding variety of reactions to the theory of evolution.

The public reception of scientific ideas depends largely on two factors: people's ability to grasp factual information and the cultural lens through which that information is filtered. The former is what scientists tend to focus on when they give popular accounts of issues such as climate change. The assumption is that if they explain things very, very clearly, everyone will understand. Unfortunately, this is an uphill battle. The general public's average capacity to weigh facts and numbers is notoriously poor — although

there is encouraging evidence that probabilistic reasoning can be improved by targeted education early in life (see page 1189).

Even more crucial, however, are the effects of the cultural lens. Over the coming month, *Nature's* Opinion pages will explore particularly vivid examples of these effects in the world's widely divergent reactions to Charles Darwin's ideas about evolution in the late nineteenth and early twentieth centuries (see page 1200).

In England, for example, the Church reacted badly to Darwin's theory, going so far as to say that to believe it was to imperil your soul. But the notion that Darwin's ideas 'killed' God and were a threat to religion was by no means the universal response in the nineteenth century.

Darwin's theory reached the world at a time when many people were looking for explanations for social, political and racial inequalities,

and in many parts of the world were wondering how to improve their lot in the face of Europe's global imperialism. So from Egypt to India, China and Japan, many religious scholars embraced Darwin's ideas, often showing how their own schools of thought had anticipated the notion of evolution. Against the threat of Western imperialism and Western charges of 'backwardness', it was to their advantage to highlight the rationality of their creed.

In China, Darwin's ideas were seen as supporting Confucians' belief in the perfectibility of the cosmic order. Evolutionary theory also became fodder for political movements of revolution and reform, and eventually laid the groundwork for communism. Latin American politicians initially reacted to Darwin's ideas by attempting to entice white Europeans to emigrate and intermarry with local populations, believing that this would 'improve the stock'. But after two world wars had made European culture look less impressive, Latin America began to see its racial diversity as an advantage, and moved towards a social view that favoured a homogeneous blend of cultures.

In nineteenth-century Russia, meanwhile, a tendency to distrust rabid, dog-eat-dog capitalism helped incline naturalists away from

a view of evolution that emphasized competition between species. Instead they embraced a 'theory of mutual aid', an account that focused on the role of cooperation in ensuring survival in a harsh environment.

The lesson for today's scientists and policy-makers is simple: they cannot assume that a public presented with 'the facts' will come to the same conclusion as themselves. They must take value systems, cultural backdrops and local knowledge gaps into account and frame their arguments accordingly. Such approaches will be crucial in facing current global challenges, from recessions to pandemics and climate change. These issues will be perceived and dealt with differently by different nations — not because they misunderstand, but because their understanding is in part locally dependent.

Darwin once said: "But then with me the horrid doubt always arises whether the convictions of man's mind, which has been developed from the mind of the lower animals, are of any value or at all trustworthy." Researchers and policy-makers would do well to mimic his humility when presenting science, and remember how people's minds truly work. ■

Mind the spin

Scientists — and their institutions — should resist the ever-present temptation to hype their results.

The circumstances surrounding the recent announcement of results from an HIV vaccine trial in Thailand are troubling.

The sponsors of the US\$119-million phase III clinical trial, a consortium led by the US Army, the National Institutes of Health and the Thai government, announced on 24 September that the trial had been a success: an analysis of the data showed that the vaccine had a statistically significant effect on preventing infection.

Other scientists could not immediately assess that claim, however: the full data from the trial were not made available until 20 October, when they were presented at an AIDS vaccine conference in Paris and in an article published online the same day (S. Rerks-Ngarm *et al.* *N. Engl. J. Med.* doi:10.1056/nejmoa0908492; 2009). The article contained two other data analyses, not mentioned in the initial announcement, showing smaller effects that were not statistically significant (see page 1187).

The trial's sponsors defend the premature announcement on the grounds that they had promised to inform the Thai people of the results first; 24 September is also Mahidol Day, the anniversary of the death of the king's father and a day of national observance in Thailand. The sponsors also argue that announcing the less-upbeat analyses along with the positive result would have been too complicated for the public to understand; they wanted to quickly deliver a clear-cut message on the trial's findings. Making the full data immediately available to scientists on 24 September would also have been impossible, they add, because of the conference and journal embargoes.

To their credit, the scientists involved did emphasize in their public statements that any vaccine effect was "modest", and that

the vaccine itself was of no immediate public-health utility. At the same time, however, they hammered home the message that this was "the first time an HIV vaccine has successfully prevented HIV infection in humans", and implied that the event was somehow historic. Such statements, together with the selective initial presentation of the data, are well outside the scientific norms for presenting the results of clinical trials. They inevitably create suspicion that the trial sponsors may have put an excessively positive spin on results that are far from clear-cut, in a trial that has long been controversial (T. V. Padma *Nature Med.* **10**, 1267; 2004). The trial has also been six years in the works, and so there seems no particular public-health urgency to justify publication by press conference.

Fortunately, such stories are still rare in science. Witness the way scientists have behaved since the beginning of the current H1N1 flu pandemic, in which the urgent threat to health creates legitimate tensions between getting results out fast and respecting peer review. Most researchers have negotiated this tension well, through a combination of fast-track publication by journals and online pre-publication sharing of preliminary data — but not through hyping their results.

Yet the temptation for scientists and their institutions to spin their research to the media, or to go publicity-mongering, is always there. And — as illustrated by the excessive public-relations campaign surrounding *Ida*, a fossil presented as a missing link in human evolution (see *Nature* **459**, 484; 2009 and **461**, 1040; 2009) — too many in the media will buy into the initial hype.

Such behaviour is corrosive to the process of scholarly scientific communication. Research institutions must not allow it to become the norm. ■

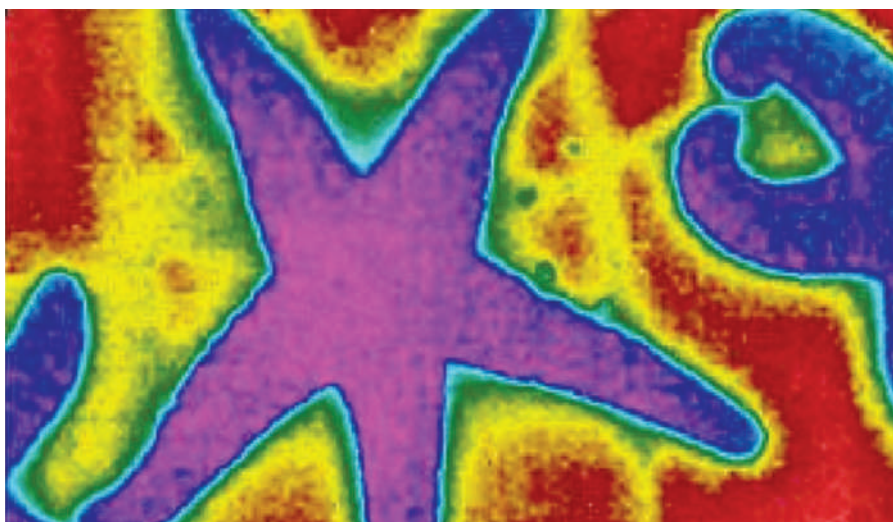
"The trial sponsors argue that announcing the less-upbeat analyses along with the positive result would have been too complicated for the public to understand."

RESEARCH HIGHLIGHTS

Chill out

Am. Nat. doi:10.1086/648065 (2009)

Sea stars know how to keep their cool when the weather heats up. By sucking up cold water while submerged at high tide, they can avoid overheating after the waves roll out. Sylvain Pincebourde, now at the University of François Rabelais in Tours, France, and his colleagues exposed ochre sea stars (*Pisaster ochraceus*) to simulated tidal cycles and various water and ambient temperatures in laboratory aquaria. They found that the intertidal predators increase the amount of colder-than-air fluid in their internal cavity after exposure to elevated aerial temperatures during low tide. This substantially reduces the sea stars' body temperatures, as shown in the infrared image, during subsequent low tides.



B. HELMUTH/S. PINCEBOURDE

PHYSICS

Quantum speed limit

Phys. Rev. Lett. **103**, 160502 (2009)

The processing speed of computer chips has doubled almost every two years for the past 40, as engineers have crammed ever more transistors into smaller circuits. But according to Lev Levitin and Tommaso Toffoli of Boston University in Massachusetts, chips will ultimately hit a roadblock, limited by the minimum time it takes for a particle to flip from one quantum state to another — a fundamental step in any information system.

There are two independent bounds on this minimum time — one based on the average energy of the quantum system, the other based on the uncertainty in the system's energy. In their calculations, Levitin and Toffoli unify the bounds and show there is an absolute limit to the number of operations that can be achieved per second by a computer system of a given energy. Levitin says that, at the current doubling pace, computing speed will reach this limit in about 80 years.

BIOLOGY

How cockroaches steer

J. Exp. Biol. **212**, 3473–3477 (2009)

Many animals can sense Earth's magnetic field, but for some species, it remains uncertain whether this ability depends on embedded magnetic particles or magnetically sensitive 'radical pair' chemical reactions of light-sensitive molecules.

Martin Vácha and his colleagues at Masaryk University in the Czech Republic exposed American cockroaches to a magnetic field in which the position of magnetic north changed by 60° every 5 minutes. This

normally makes cockroaches restless. The team also applied a radio-frequency field at only a fraction of Earth's field intensity to jam the creature's magnetic sensing system. At a certain frequency, the cockroaches stayed calm. Other radio frequencies had the same effect, but at higher field strengths.

Because the radio-frequency field should not affect a magnetic-particle-based sensor, the result suggests that insects use a radical-pair-based method of sensing magnetic fields.

NANOTOXICOLOGY

Lung penetration

Nature Nanotechnol. doi:10.1038/nnano.2009.305 (2009)

When inhaled by mice, multiwalled carbon nanotubes (CNTs) can embed themselves in the lining of the lung (pictured below).

James Bonner at North Carolina State University in Raleigh and his colleagues exposed mice to nanoparticle aerosols of either 30 milligrams per cubic metre or 1 milligram per cubic metre for six hours. In the mice exposed to the higher level, immune cells called macrophages (pictured) engulfed the nanotubes and carried them to the lung lining.

Within weeks of exposure, those mice also developed a condition called subpleural

fibrosis, which causes localized fibrous lesions. This work does not confirm the suggestion made by other studies that nanotubes may cause lung tumours, but the authors say they urge caution be taken when people are exposed to nanotubes in the air.

CLIMATE CHANGE

Stormy warming weather

Geophys. Res. Lett. doi:10.1029/2009GL039810 (2009)

Arctic residents have complained that their weather is changing, and a study by researchers in Australia backs this up.

Ian Simmonds and Kevin Keay of the University of Melbourne examined the relationship between the number and strength of cyclones each year in the Arctic basin and the extent of Arctic sea ice during the month of September for the past 30 years. They found that years with the least amount of ice had significantly stronger storms, although there was no correlation with the number of storms.

The findings support previous forecasts that the decline in sea ice, and particularly the record lows of recent years, is raising the risk of stronger storms in the Arctic.

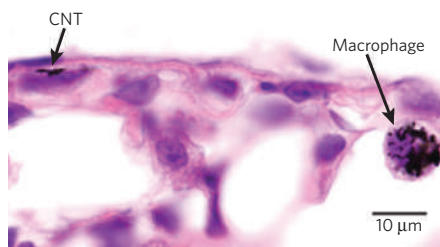
CANCER BIOLOGY

Double protection

Sci. Trans. Med. **1**, 3ra7 (2009)

Blocking a key cell-signalling pathway increases certain tumours' sensitivity to high doses of ionizing radiation while protecting healthy tissue from the harmful rays.

David Roberts of the National Cancer Institute in Bethesda, Maryland, and his colleagues inhibited the cell-surface receptor



CD47 or the protein that binds it, TSP1, which regulates cell growth and survival in response to stress, such as that caused by radiation. They found that suppressing the CD47–TSP1 pathway in normal human cells improved their survival after irradiation and, in mice, led to reduced radiation injury.

In addition, the tumours of mice treated with a CD47-blocking molecule prior to radiation exposure were up to 89% smaller 30 days after irradiation than those of mice receiving radiation alone.

BIOPHYSICS

All seeing eye

Nature Photon. doi:10.1038/nphoton.2009.189 (2009) Polarized light is used in optical devices, including some microscopes. Being able to control polarized light is key. Materials such as crystals can do the job, but only within a limited range of wavelengths.

Nicholas Roberts of the University of Bristol, UK, and his colleagues have worked out how a species of mantis shrimp can switch polarized light from one form to another over a range of colours. A thin band of specialized receptor cells in the eyes of *Odontodactylus scyllarus* have just the right structure, dimensions and composition to enable them to control polarization over most of the visible spectrum. The team believes that further study of this mantis shrimp's eyes could lead to better optical devices.

NEUROSCIENCE

Brain signal source

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0905509106 (2009)

Functional magnetic resonance imaging (fMRI), used to map brain activity, gives a signal when the levels of oxygenated blood increase. The signal is often preceded by a darkening, thought to indicate early oxygen absorption by the brain owing to local neural activity.

But a study by Aniruddha Das and his colleagues at Columbia University in New York casts doubt on this. They used intrinsic signal optical imaging, a technique similar to fMRI, to measure changes in blood volume and blood oxygenation in the brains of two macaques while they performed a visual task.

The researchers found that during the initial darkening, blood-oxygen levels changed little, but blood volume increased markedly. The team suggests that blood-volume change is a better signal to use in brain imaging because it seems to be more closely linked to neural activity, occurring even before changes in blood oxygenation.

SEXUAL SELECTION

Intruder alert!

Proc. R. Soc. B doi:10.1098/rspb.2009.1554 (2009)

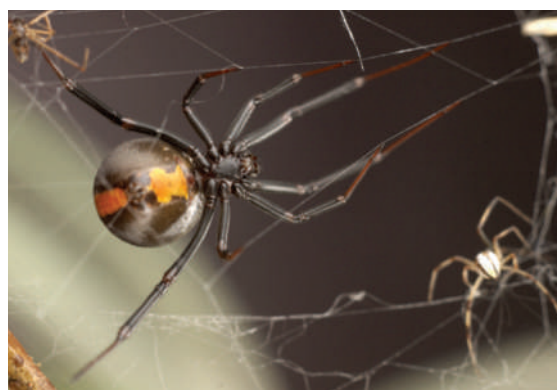
Male redback spiders can sneak in and quickly copulate with a female after a rival male has already spent hours wooing her, yet avoid the usual penalty of short courtship — being eaten prematurely by his lover.

The female Australian redback spider (*Latrodectus hasselti*, pictured below) eats the male after mating, but sometimes consumes him prematurely — after he's copulated only once. Jeffrey Stoltz and Maydianne Andrade of the University of Toronto in Canada measured the spiders' courtship durations and found that females tend to eat their partners prematurely if courtship is less than 100 minutes long. However, intruder males can mate after a shorter courtship and avoid premature death if an earlier male had already exceeded this 100-minute threshold.

This could lead to lower quality males seeking out, rather than avoiding, competition with rival spider studs, the authors say.

For a longer story on this research, see <http://go.nature.com/U6DPEG>

K. JONES



ASTRONOMY

Galaxy size matters

Astrophys. J. **705**, 255–260 (2009)

A survey of distant galaxies shows that more loosely packed ones tend to form more stars.

The survey looked at 225 galaxies at distances of between about 2.8 and 3.4 parsecs from Earth. It found that compact galaxies tend to have fewer new stars than do their larger counterparts of comparative mass.

Sune Toft of the University of Copenhagen and his colleagues conclude that compact galaxies formed many stars quickly in one intense burst, early in the history of the Universe. Conversely, larger, more diffuse galaxies form stars gradually over a longer period of time. The results may explain why very distant galaxies are often more compact than the younger ones nearby.

JOURNAL CLUB

Jonathan Weissman
University of California, San Francisco

A biochemist looks at how DNA sequencing can reveal more than just sequences.

Huge advances in DNA sequencing have allowed us to readily determine the sequence of almost any living (and a few extinct) species. Yet arguably, most biological insight comes from work on five model organisms: *Escherichia coli*, baker's yeast, roundworms, fruitflies and mice. Unfortunately, many important biological processes are not captured in these creatures.

Papers from two groups, one led by Andrew Camilli of Tufts University in Boston, Massachusetts, the other by Brian Akerley at the University of Massachusetts in Worcester, describe new genetic tools that allow the quantitative dissection of gene function in a wide range of microorganisms (T. van Opijnen *et al. Nature Methods* **6**, 767–772; 2009; and J. D. Gawronski *et al. Proc. Natl Acad. Sci. USA* **106**, 16422–16427; 2009). These studies combine exhaustive transposon mutagenesis — whereby thousands of small DNA segments, or transposons, are introduced into the genome to mutate many genes — with massively parallel, or 'deep' sequencing of transposon/chromosome junctions to monitor the consequences of the loss of single or pairs of genes on the organisms' traits.

The real power of the approaches comes from the deep sequencing, which tracks the abundance of individual transposon mutants after they have been subjected to a stress. Knowing by how much each mutant has grown or suffered under the stress provides a measure of the relative roles that the mutated genes have.

I find it particularly gratifying that the advances in deep sequencing that have allowed us to catalogue so many genes from so many organisms can now be harnessed to help us figure out what these genes actually do.

Discuss this paper at <http://blogs.nature.com/nature/journalclub>

NEWS BRIEFING

● POLICY

Space-flight review: The US panel charged with reviewing NASA's human space-flight programme issued its final report last week, and warned that the current programme seems to be "on an unsustainable trajectory". Present funding doesn't match the space agency's targets, says the commission, which is chaired by ex-aerospace executive Norman Augustine. Many of its suggestions, such as bypassing human exploration of the Moon and scrapping the Ares-I rocket in favour of commercial space flights, had already been aired in public meetings (see *Nature* 460, 791; 2009).

European research reform: The European Commission has agreed that a top scientist should lead the administrative and managerial activities of the European Research Council (ERC), in place of the commission's current appointee, economist Andreu Mas-Colell. The decision, announced on 22 October, came in response to a damning review of the ERC published in July (see *Nature* 460, 557; 2009), which called for immediate reforms to the council's management. Fotis Kafatos, president of the ERC, said that the commission's response was welcome but "not particularly revolutionary". See go.nature.com/Eh8n43 for more.

Nuclear vision: Germany's new coalition government will extend the lifespan of the nation's nuclear power plants — which last year produced around 23% of the country's electricity needs — beyond 2022. But the Christian Democratic Union and its junior coalition partner, the liberal Free Democratic Party, will not revise an existing ban on building new nuclear plants. In a 24 October policy plan, the coalition also agreed to "immediately" lift a moratorium on evaluating the merits of the Gorleben salt dome, a controversial storage site for nuclear waste.



AHN YOUNG-JOON/AP

HWANG CONVICTED

Disgraced South Korean cloning scientist Woo Suk Hwang left Seoul Central District Court on 26 October knowing that his sentence, a two-year prison term suspended for three years, could have been worse. He was found guilty of embezzling government funds and buying human eggs in violation of the country's bioethics law, but was cleared of fraud. The prosecution had sought a four-year jail term, and plans to appeal. See page 1181 for more.

Mercury deadline: The US Environmental Protection Agency (EPA) has agreed to set new rules governing emissions of mercury and other toxic chemicals from power plants by November 2011, according to a settlement in a federal lawsuit filed by several environmental and health groups. Environmentalists say that the Clean Air Act required the EPA to set limits by 2002, but the administration of former US President George W. Bush avoided this in part by creating a market-based system that would allow mercury emissions to continue at some plants as long as they dropped in aggregate. However, Bush's work-around was later deemed illegal in federal court.

Polar-bear protection: The US Fish and Wildlife Service proposed on 22 October to designate around 500,000 square kilometres of "critical habitat" — 96% of which is sea ice — for

the polar bear. The bear was listed as a threatened species in 2008 owing to projections of sea-ice declines caused by global warming. The government is already obligated to avoid actions that jeopardize the bear, but the designation would add another layer of protection by also making it illegal to conduct activities that adversely affect the bear's habitat.

NUMBER CRUNCH

57%

of Americans think there is "solid evidence the Earth is warming", according to an October 2009 poll.

71%

of Americans thought this in April 2008.

Source: Polls conducted by Washington-based Pew Research Center for the People & the Press

Vaccine report: More children than ever are being immunized, but 24 million infants in the world's poorest nations still do not receive routine immunization, according to a report by the World Health Organization, UNICEF and the World Bank. The 21 October *State of the World's Vaccines and Immunization* report says that although four in five children now have access to life-saving vaccines, at least another US\$1 billion is needed annually to help raise immunization rates above 90%. This would cover

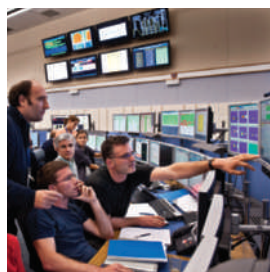
the rising costs of immunization and could prevent two million children from dying per year by 2015. See go.nature.com/APMPtB for more.

GM protests: Environmental groups are protesting after the Mexican government's 15 October approval of the first permits to plant experimental genetically modified (GM) maize (corn). Growing of GM varieties had previously been outlawed in the country, which is the homeland of domesticated maize. Mexican environmental and agricultural agencies say that they will keep plantings away from traditional 'landrace' maize, and will monitor the experimental crops closely before considering requests from agribusiness for full-scale GM maize planting. But researchers say that past landrace contaminations from illegal GM maize planting (see *Nature*, 456, 149; 2008) mean corruption of traditional genomes is inevitable.

● FUNDING

Energy funding: The US Department of Energy has awarded \$151 million to 37 research projects through the recently formed Advanced Research Projects Agency-Energy (ARPA-E). Based on the Defense Advanced Research Projects Agency, ARPA-E is geared towards high-risk 'transformational' energy research that might not be funded through traditional science grants. Awardees included small businesses, educational institutions and

NEWS MAKER



Large Hadron Collider

Physicists last week injected particles into the accelerator for the first time since an accident forced it to shut down in September 2008.

large corporations that focus on everything from liquid-metal batteries and gasoline-producing bacterial reactors to new methods for making light-emitting diodes and synthetic enzymes for capturing carbon dioxide from industrial emissions.

Innovation fund: The United States is following up on promises to facilitate a global fund to trigger innovation and technology development. On

23 October the government's Overseas Private Investment Corporation issued a call for proposals for a Global Technology and Innovation Fund, aimed at countries in Asia, the Middle East and Africa. Each selected proposal would receive between \$25 million and \$150 million; potential areas include clean technology and information technology.

● RESEARCH

Scientific espionage: A former Los Alamos nuclear-weapons physicist says that he is under investigation for espionage. The researcher, P. Leonardo Mascheroni, spoke to the Associated Press on 22 October, two days after he says FBI agents raided his home. The bureau confirmed an "ongoing investigation" into his activities. Mascheroni, who worked in the lab's X Division in the 1980s, says that last year he gave unclassified information — widely available on the Internet — to a man claiming to be from the Venezuelan government who asked for information about starting a nuclear-weapons programme.

HIV vaccine doubt: Results of the largest-ever HIV-vaccine trial looked less impressive when full details were formally published last week (S. Rerks-Ngarm *et al.* *N. Engl. J. Med.* doi:10.1056/nejmoa0908492; 2009) than when they were outlined in a press release a month earlier. In September, the trial was said to show that a vaccine combination reduced the risk of HIV infection

THE WEEK AHEAD

29 OCTOBER-1 NOVEMBER

Philadelphia hosts the 47th Annual Meeting of the Infectious Diseases Society of America.

► go.nature.com/ykfvNW

29-30 OCTOBER

A European Council summit meeting in Brussels may firm up European promises to finance climate-change action in developing countries.

► go.nature.com/1kWXLS

2 NOVEMBER

The European Space Agency is scheduled to launch its Soil Moisture and Ocean Salinity satellite.

► go.nature.com/sHQ161

2-6 NOVEMBER

Nairobi, Kenya, hosts the Multilateral Initiative on Malaria's fifth Pan-African Malaria Conference.

► www.mimalaria.org/pamc

2-6 NOVEMBER

The United Nations Framework Convention on Climate Change holds its fifth round of international climate talks this year in Barcelona, Spain.

► go.nature.com/QsS4jx

by nearly one-third. But Peter Smith, a tropical epidemiologist at the London School of Hygiene & Tropical Medicine, says "there is not much evidence from the data that it protects at all". See page 1187 for more.

BUSINESS WATCH

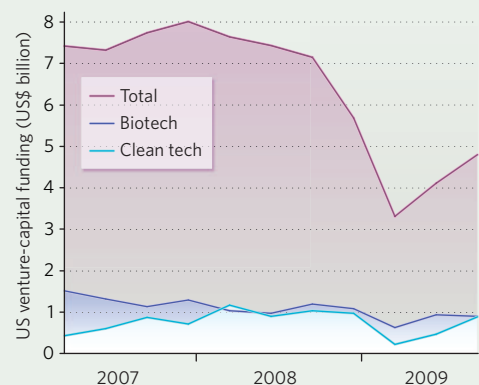
Beginning its rehabilitation after a dismal funding scene earlier this year, venture-capital financing in the United States saw a small improvement in the year's third quarter (to 30 September). It increased by 17% over the previous quarter to \$4.8 billion, according to a 20 October MoneyTree Report by PriceWaterhouseCoopers and the National Venture Capital Association, based on data from Thomson Reuters.

The clean-technology sector, which spans industries such as alternative energy, conservation, pollution-scrubbing, recycling and power supply, was responsible for more than half of the rise. Venture-capital

funding in this sector grew by 89% from the previous quarter to \$898 million in 57 deals. A few large funding rounds fuelled this increase; three of the top ten deals of the third quarter went to clean-tech companies in California: \$286 million to Solyndra of Fremont (photovoltaics), \$82.5 million to Tesla Motors of San Carlos (electric vehicles) and \$60 million to Serious Materials in Sunnyvale (energy-efficient building materials).

Biotechnology continues to be the top funded sector, receiving \$905 million in the third quarter — a 4% decrease from the second quarter.

CLEAN TECH AIDS FUNDING REHAB



SOURCE: PWC

NEWS

African science feels the pinch

Recession dampens donors' enthusiasm.

DURBAN

The global financial crisis is hampering plans to revive African science, researchers and policy-makers said last week in Durban, South Africa. Slashed donor funding, slowing foreign investment and competing budget priorities are the main culprits; hardest hit are the poorest countries and continent-wide projects.

"Our countries are in crisis, philanthropists are in crisis and the aid agencies are in crisis," Jean-Pierre Ezin, commissioner for science, technology and human resources for the African Union, said at a conference organized by TWAS, the academy of sciences for the developing world, based in Trieste, Italy.

In 2007, an African presidential summit on science saw funders falling over each other to offer assistance on science and technology programmes. Today the funding situation has changed dramatically.

The Swedish international development agency SIDA said last month that it would cut funding for its research cooperation programmes with developing countries by 20%, from an estimated 1.05 billion kronor (US\$150 million) to 800 million kronor. Britain's Wellcome Trust, which funds several medical research projects in Africa, cut overall grants for 2008–09 by £30 million (US\$50 million), to £590 million. Many people expect aid levels to fall this year as a result of the financial crunch, although the Organisation for Economic Co-operation and Development says effects are more likely to be felt in future years, as 2009 aid budgets were mostly finalized before the recession hit.

Not all the news is bad. Several American philanthropists, including the Bill & Melinda Gates Foundation, said earlier this year that they would not cut research funding. The Gates Foundation even said that it would increase grants, despite a 20% drop in assets last year.

Still, Ezin says that the African Union won't be able to fulfil all its planned science activities for 2009 and 2010. Instead, his department will prioritize the Pan-African University, a network of existing African institutions that will train PhDs and carry out research, and a grant programme for African researchers. But a programme to train science teachers, he says, may have to be scaled back.

University researchers are also feeling the



R. FREMSON/NEW YORK TIMES/REDUX/EVEVINE

Broken hopes: research projects between African universities and other countries are being cut back.

pinch. In Senegal, a plan to expand the country's university system has come to a standstill. "The crisis came and everything stopped," says Lamine Ndiaye, a former vice-chancellor of the University of Gaston Berger in Saint-Louis, Senegal. Funding has dried up from the government and from France, the country's main development partner, he says.

Countries that don't depend on aid are also struggling. In Nigeria, the drop in demand for oil and gas, exacerbated by a stricken banking sector, means that private donations — a major source of funding for Nigerian universities — are slowing. "In the past, a conference like this would have a lot of Nigerians coming, supported by industry grants. We don't find many today," says Oye Ibidapo-Obe, president of the Nigerian Academy of Science in Lagos.

Nigeria's government won't pick up the slack left by the drop in private investments, Ibidapo-Obe adds. "Research is not seen as the major driver of the economy."

Even South Africa, the continent's economic powerhouse, is facing a lean year. The country's coffers have been depleted by its worst financial performance since the end of apartheid 15 years ago, says Naledi Pandor, the science minister.

Pandor says she has been assured by the country's treasury that her department won't face cuts in the mid-term budget, due for release as *Nature* went to press. But the dip in the country's growth rate means that the department, which was given 4.2 billion rand (US\$560 million) for 2009–10, may have to put some planned projects on the back burner. Probable cuts include a 700-million-rand semi-commercial titanium test facility, which could be either delayed or dropped completely.

Recent progress in building up African science could be lost, warns Mohammed Hassan, executive director of TWAS. In the 1980s, African governments responded to a steep economic decline by cutting their higher-education budgets, he notes, and Africa went from having some of the best universities in the developing world to some of the worst.

Although the world economy is starting to recover, the worst may be yet to come for Africa, says John Muyonga, a food scientist at the University of Makerere in Uganda. His institution depends "close to 100%" on donor funding, he says. Most of his colleagues have grants that cover several years, and may struggle to find new funding when these grants run out. "We may have more impact in 2010 than in 2009," he adds.

Linda Nordling





Q&A: A HEALTHY VISION
Tom Frieden, director of the US Centers for Disease Control and Prevention, speaks to *Nature*.
go.nature.com/Py9Srb

CDC

Woo Suk Hwang convicted, but not of fraud

Cloning pioneer Woo Suk Hwang was sentenced to two years in prison at the Seoul Central District Court on 26 October, after being found guilty of embezzlement and bioethical violations but cleared of fraud.

Supporters of Hwang, a former professor at Seoul National University in South Korea, were pleased with the sentence, which is suspended for three years and half the length sought by prosecutors. The prosecution plans to appeal.

Hwang was once fêted for creating human stem-cell lines using cloned embryos derived from patients suffering from spinal-cord injury and other disorders (W. S. Hwang *et al.* *Science* 303, 1669–1674; 2004 and W. S. Hwang *et al.*

Science 308, 1777–1783; 2005). The accomplishment, which promised an endless supply of stem cells genetically matched to patients, turned out to be bogus.

Hwang admitted in January 2006 to falsifying data, while maintaining that he had the ability to do what he had claimed. In South Korea, scientific fraud would be illegal only if Hwang had used fraudulent data to gain grants. Prosecutors argued that he duped two companies, SK Group and NongHyup, into supplying research funds. But according to media reports, the court rejected the allegations on the grounds that the firms provided money without expecting to benefit.

The court did, however, find

Hwang guilty of buying human eggs in violation of the country's bioethics law and of embezzling 830 million won (US\$700,000) of government money.

The Korea Times reported that the light sentence was motivated by judge Ki-ryul Bae's sympathy for Hwang's apparent dedication to Korean biotechnology and his stated remorse. Hwang will now be able to focus on his research career, which he has been rebuilding since he was indicted in May 2006 (see *Nature* 461, 1035; 2009).

Many researchers are not ready to welcome Hwang back. "It was not just one moment of weakness — the degree of manipulation of the goodwill of people, particularly

fellow scientists, made it more," says Alan Colman, a stem-cell scientist at the Institute of Medical Biology in Singapore. "The sad thing is that it's clear he is a talented experimentalist." Colman argues that Hwang should not be eligible for research funding from public sources for a prolonged period.

Researcher Ryuzo Torii of the Shiga University of Medical Science in Japan used large amounts of grant money, time and monkey eggs trying to reproduce Hwang's technique in non-human primates in 2004 and 2005. He says that forgiving Hwang and recognizing him as a researcher would be "a mistake".

David Cyranoski

US physicists propose astrophysics goals

Cosmic-ray experiments would suffer, unless budgets increased, under a list of priorities for high-energy physics given to the US Department of Energy (DOE) by independent advisers last week.

The DOE should instead fund one massive dark-matter detector, one major dark-energy experiment and a high-energy gamma-ray detector, according to the report, which was presented on 23 October in Washington DC.

For current budget trends, the report did not endorse Auger North, a US\$127-million array of 4,400 cosmic-ray detectors in southeastern Colorado. This is due partly to the success of its southern counterpart, the Pierre Auger Observatory in Argentina, which is close to tracing the origin of high-energy cosmic rays to massive black holes in galactic centres, says Steve Ritz, chair of the 15-person committee that wrote the report. "You no longer need to invoke new physics to explain [high-energy cosmic rays]," says Ritz, an astrophysicist at the University of California, Santa Cruz.

The report is the first in which US high-energy physicists have set priorities for particle astrophysics. The committee considered four budget scenarios, ranging from the relatively grim DOE budget for the 2008 fiscal year in years going forwards,



Southern cosmic-ray observatory may not get a partner.

to budgets that would more than double over the next decade. The DOE takes the advice "very seriously," says Dennis Kovar, the agency's associate director of science for high-energy physics.

Even in the tightest scenario, dark-matter experiments fared well — both because they are relatively cheap and because there are hints that discovery of these undetected particles, which make up nearly a quarter of the Universe's mass-energy, could be imminent. If budget trends continue, however, the report says that only one next-generation detector can be chosen from among competing groups, which propose

ten-tonne vats of liquid argon or xenon, or cryogenic silicon detectors set deep underground.

Dark energy — the mysterious vacuum force that makes up most of the rest of the Universe's mass-energy — remains a priority, although the experiments are expensive (see page 1182). Unless budgets increase, the report suggests, the DOE will have to choose between supporting a space-based project, the Joint Dark Energy Mission, or a major ground-based telescope: either the 8.4-metre Large Synoptic Survey Telescope (LSST) or BigBOSS, a 4-metre telescope.

"We don't have the luxury of a lot of redundancy," says Patricia Burchat, a physicist at Stanford University in California and a member of the LSST project. "We can't afford it." Ritz says that so far, the agency hasn't looked at the trade-offs between ground- and space-based approaches to dark energy.

For high-energy gamma-ray astrophysics, the report supports the Advanced Gamma-ray Imaging System, an array of about 50 telescopes spread over a square kilometre — but only if the consortium merges with a similar European proposal, the Cherenkov Telescope Array.

Eric Hand

PIERRE AUGER OBSERVATORY



Measuring supernovae, such as this one (bright spot, bottom) in the galaxy NGC 4526, could help pin down dark energy.

Dark energy rips cosmos and agencies

An international space mission to study an astronomical mystery is foundering.

A once-favoured space probe to study dark energy is struggling to get off the ground, as three agencies in the United States and Europe tussle over the details of a potential international mission.

The rise and fall this year of the Joint Dark Energy Mission (JDEM) — a satellite meant to pin down the repulsive force that is accelerating the Universe's expansion — is partly due to strife between two US agencies, NASA and the Department of Energy (DOE), and a third potential partner, the European Space Agency (ESA). In addition, scientists working on the JDEM designs have not presented a unified front, owing to disagreements over the best observational method to use (see 'Hunting for dark energy') — at a time when an influential astrophysics panel is about to prioritize the next decade's best and most organized missions.

"This is an example of a satellite blowing up before it gets built," says Bob Nichol, an astrophysicist at the University of Portsmouth, UK, who is working on the European design concept.

Dark energy is a fudge factor for a force that has the upper hand on gravity, pushing the Universe to accelerate at an ever-faster rate. The effect was announced in 1998 after

astronomers precisely measured the distances to supernovae in other galaxies. But the cause remains baffling. "It is perhaps the biggest mystery of our time," says Neil Gehrels, JDEM project scientist at the Goddard Space Flight Center in Greenbelt, Maryland. "It determines the fate of the Universe." Depending on how strong it turns out to be, dark energy could dissipate the Universe into darkness, shred it in a 'big rip' or even reverse itself, allying with gravity to create a 'big crunch'.

JDEM was meant to figure this out, and the mission had early momentum. It got a political boost from a 2007 study by the National Research Council that ranked a dark-energy probe as the top priority for studying deep cosmological questions. Then, to forge a scientific consensus, in September 2008 NASA and the DOE brought together three design teams that had been competing to fashion a mission that could potentially accommodate all of them (see *Nature* 455, 577; 2008).

Hunting for dark energy

Scientists are arguing over three space-based approaches to measure dark energy.

Supernovae Measuring distances to supernovae was the first, and best-understood, way to probe dark energy, but also has the most limitations. It assumes that all supernovae in the early Universe exploded with the same characteristic brightness.

Baryon acoustic oscillations

Sound waves from the Big Bang imprinted ripples in the distribution of galaxies. By comparing the size of the ripples in early galaxy clusters with those in clusters that formed later in the Universe, astronomers can deduce the effect of dark energy. But finding enough young clusters big enough to work with is a challenge.

Weak lensing

This looks for tiny distortions in the shapes of galaxies caused by intervening dark matter, which varies with cosmic time and relates to dark energy. It requires images of many galaxies from surveys, which means that such a mission would need to operate differently from one optimized for baryon acoustic oscillations.

E.H.

NASA, ESA, HUBBLE KEY PROJECT TEAM, HIGH-Z SUPERNOVA SEARCH TEAM

Two months later, the agencies inked an agreement in which NASA would lead the mission. The DOE has said it would contribute roughly one-quarter of the cost.

But price was a problem. Separately, each of the three teams had missions with price tags of more than US\$1 billion; combined, the JDEM was just as expensive, if not more so. NASA's astrophysics division had always wanted something smaller and cheaper. But "how much dark energy can you buy for \$600 million?" asked Jon Morse, director of NASA's astrophysics division, at an advisory meeting this month. "No one seems to want to answer that question."

So NASA went to ESA, which had been pursuing its approach, called Euclid, for years. By January, it looked as if ESA — and its budgetary resources — were on board (see *Nature* doi:10.1038/news.2009.38; 2009).

Downgraded

But when NASA told DOE officials that the Europeans would be building scientific instruments that the DOE had planned to make itself, the energy agency balked. The DOE–NASA agreement specifies that the DOE would build a "major scientific instrument". Yet in negotiations through the winter, the department was offered minor-league work that some there considered insulting.

"I don't disagree that there were unhappy scientists on all three sides," says Gehrels. But no division of duties was set in stone then, he says. And he points out that very few multi-agency missions are smooth; for instance, the \$700-million Fermi Gamma-ray Space Telescope, with contributions from the DOE and NASA, went through a difficult birthing process for more than a decade before being launched last year. At a physics advisory meeting last week in Washington DC, William Brinkman, the DOE's director for the Office of Science, said of JDEM: "It's not an uncontroversial situation."

DOE officials and scientists were so unhappy that, for a few months this spring, they broke off from the mission. Some, at the Lawrence Berkeley National Laboratory in California, instead proposed a ground-based dark-energy survey called BigBOSS. In presentations in June, the team claimed that BigBOSS would rival JDEM's results for an overall cost of just \$85 million.

Meanwhile, NASA and ESA ploughed ahead with a design that became known as the International Dark Energy Cosmology Survey (IDECS). In April, this partnership also ended because the merger wasn't happening quickly enough. In Europe, the Euclid team was told to resume work on its 1.2-metre space telescope, Nichol

says, in preparation for a triage in February 2010 under ESA's 'Cosmic Visions' competition.

On the US side, three design concepts remain in play. The \$1.6-billion IDECS design has two types of detectors and would use all three observational methods on a 1.5-metre telescope. A \$1.2-billion 'Omega' design would use only one detector and sacrifice some capabilities. Gehrels says that his team has now been tasked to look at a third concept, with a price cap of \$850 million, that will require a smaller telescope and cut the probe's capabilities.

Meanwhile, as the space-based mission ideas stall, ground-based astronomers are making strides in pinning down uncertainties in a key dark-energy parameter that determines whether the force is a constant or changes with time. One of many recent studies, which added more than 100 newly observed supernovae to existing studies, suggests that it is a constant with a precision of 7% (M. Hicken *et al. Astrophys. J.* **700**, 1097–1140; 2009).

Simon White, director of the Max Planck Institute for Astrophysics in Garching, Germany, questions whether it is worthwhile to spend a billion dollars just to show, with more precision, that dark energy is a constant. "There's nothing to look for," says White. "It's fairly close to a cosmological constant, and the question [for

"This is an example of a satellite blowing up before it gets built."

JDEM] is: is it very, very close to a cosmological constant?"

But all the recent studies that seem to be honing in on a cosmological constant make the assumption that it is constant in time, says Rocky Kolb, a theorist at the University of Chicago in Illinois.

To achieve precision and to test how dark energy has varied in time, Kolb says, going to space is necessary. Both possibilities — a constant, or a changing dark energy — are equally ugly to theorists, who have few explanations for either scenario.

Kolb notes that there is a third possible result. Several ground-based microwave telescopes, such as the South Pole Telescope, are tracking how the structure of very distant galaxy clusters grew in the early Universe under the influence of gravity. If these results, expected soon, do not agree with the measurements of the expansion history of space — the measurements made by methods used by JDEM — it would indicate that something is wrong with general relativity.

But JDEM may not get to make these comparisons. NASA and the DOE will not commit to any of the designs until after the community survey weighs in in spring 2010. "We'll wait for the decadal survey," Morse said earlier this month, "to see if we have a mission." ■

Eric Hand

Ozone protocol squares up to climate

Europeans back efforts to amend the Montreal Protocol to address global warming.

As climate observers fret over December's global-warming summit in Copenhagen, international ozone negotiators are quietly plugging away on a proposed amendment to the Montreal Protocol on ozone-depleting substances — one that could reshape the way governments tackle an important class of greenhouse gas.

Delegates to the protocol will meet in Port Ghalib, Egypt, on 4–8 November to discuss proposals to reduce the use of hydrofluorocarbons (HFCs), common refrigerants that were deployed to replace ozone-destroying chemicals such as chlorofluorocarbons. Although HFCs do not damage the ozone layer, some are up to 12,000 times more effective than carbon dioxide at trapping heat. Advocates of the proposals say that HFCs could be efficiently regulated under the Montreal Protocol — which has proved effective in cutting the use of ozone-depleting substances — to help slow global warming.

European Union (EU) negotiators endorsed the idea during the United Nations climate talks in Bangkok earlier this month, adding momentum to a movement led by the small island states of Mauritius and Micronesia, which are seeking aggressive action on climate change to stave off rising sea levels. Those states submitted an initial amendment to the protocol in April, followed last month by a proposal from the United States, Canada and Mexico.

Nonetheless, it is unclear whether Montreal delegates will move forward with such a decision when they meet in Egypt. European officials remain wary of letting the Montreal process get



A real air con: local cooling, but global warming.

too far ahead of the Copenhagen climate talks (see 'On the global-warming front'), and the idea has yet to garner endorsement from emerging economies such as China and India, which represent the primary growth market for HFCs owing to rising demand for air conditioning.

"I could envision something that says, 'We are willing and interested in the possibility

and would be prepared to take this on if so instructed,'" says Ana Maria Kley Meyer, Argentina's former lead negotiator on the protocol, who now works at the International Centre for Trade and Sustainable Development in Geneva, Switzerland.

Montreal delegates first ventured into global warming in 2007 when they agreed to accelerate the phase-out of hydrochlorofluorocarbons to reduce the gases' greenhouse effects. Regulating HFCs, however, would require an explicit expansion of the Montreal treaty.

Language proposed by European climate negotiators would allow their Montreal counterparts to move forward on HFCs as long as any greenhouse-gas reductions achieved by decreasing HFCs could count towards commitments under the UN climate treaty.

"The Europeans just want to make sure that whatever is done on the Montreal Protocol side is credited on the climate side," says Durwood Zaelke, president of the Institute for Governance & Sustainable Development in Washington DC. He says one possibility is that EU negotiators in Egypt could agree to set global HFC emission standards for vehicles, consistent with current EU domestic regulations.

The European Commission says it wants to resolve how to coordinate financing, accounting and regulations between the Kyoto and Montreal treaties before moving forward. The commission is calling on a technical committee established under Montreal to analyse the proposals by the middle of 2010.

Jeff Tollefson

On the global-warming front

As ozone negotiators look for ways to help battle global warming (see above), the international representatives who work solely on climate are inching along towards their own deal.

Last week in Luxembourg, European Union (EU) environment ministers fleshed out their negotiating position for the climate summit in Copenhagen this December, but finance ministers were unable to agree on how much money the EU will

commit to developing countries for adaptation and mitigation to climate change.

The environment ministers endorsed earlier commitments to cut emissions by 20% by 2020 compared to 1990 levels, and by 30% if there is a global treaty; they also said that emissions from the aviation and maritime sectors should be curbed by 10% and 20%,



respectively, below 2005 levels by 2020.

European heads of state are expected to approve these recommendations and to tackle the issue of climate financing when they meet in Brussels this week.

The European Commission has estimated that, in total, developing countries will need €100 billion (US\$150 billion) annually to begin

developing low-carbon economies to cope with the impacts of climate change.

In the United States, meanwhile, Senate Democrats have begun work on a bill to curb US emissions by 20% by 2020 and by 83% by 2050, compared with 2005 levels. The Senate Environment and Public Works Committee is hearing testimony on the bill this week, to be followed "as soon as possible" by amendments and a final vote.

J.T.



FATAL FROG FUNGAL DISEASE FIGURED OUT
Electrolyte imbalance stops amphibians' hearts.
go.nature.com/BOUwd

V. T. VREDENBURG/SFSU

University tightens oversight of sensitive research

University administrators are looking to sharpen their monitoring of export violations, officials said last week at a meeting of the National Council of University Research Administrators in Washington DC.

The move comes in the wake of the first US conviction, last year, of a university professor for trafficking military-sensitive information. In July, John Reece Roth, formerly an engineer at the University of Tennessee, Knoxville, was sentenced to four years in prison for breaching the Arms Export Control Act; he remains free pending an appeal. Roth had shared sensitive information relating to a plasma-guidance system for unmanned aircraft with a graduate student from Iran and another from China. The case has triggered anxiety among many academics, who fear that they could be punished for unintentional slips (see *Nature* 461, 156; 2009).

"Now that the faculty members know the facts of the Roth case, they don't want to be individually challenged; they want the support of the university," says David Brady, director of export and secure research compliance at Virginia Polytechnic Institute in Blacksburg.

Roth's project "slipped through the cracks" of the university's monitoring system, says Robin Witherspoon, export-control officer for the University of Tennessee's office of compliance. Witherspoon says that the university now has electronic flagging systems in place to alert her office to suspicious financial- or travel-related dealings with other countries, as well as any potentially problematic grant proposals or contracts. Witherspoon has also instigated training programmes to educate researchers

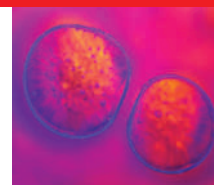
about export control of sensitive data or technology.

Christopher Golomb, a special agent with the FBI counterintelligence division in Washington DC, says that in light of the Roth conviction, the bureau is also adjusting the ways that it interacts with academic institutions. Last year, the agency conducted a survey with the Federation of American Scientists to assess negative views of law enforcement held by some in the scientific community. The

FBI is also engaged in two academic alliances with university and college presidents. "It's a great lesson learned," says Golomb. "It helps us develop policies and procedures so that professors know that they can be targeted."

The White House has ordered a review of current US export control regulations. ■
Elie Dolgin

"It helps us develop policies and procedures so that professors know they can be targeted."



Q&A: MAKING A CELLULAR MENAGERIE
Caroline Kane talks about a new image library of the cell.
go.nature.com/N9duow

GETTY

Jury still out on HIV vaccine results

Some experts see hope in trial findings, but others say that the data do not back up such optimism.

PARIS

More than 1,000 researchers in Paris last week rapturously applauded formal results from the largest-ever HIV vaccine trial. In a preliminary announcement in September, the trial, which included 16,402 people in Thailand, was said to show that a vaccine combination reduced the risk of HIV infection by nearly one-third (see *Nature* doi:10.1038/news.2009.947; 2009).

But some scientists are sceptical, arguing that the response of the HIV research community, long deprived of any good news from vaccine trials, is based more on hope than on rigorous science.

The US\$119-million phase III trial, sponsored by the health ministry of Thailand and the US Army, started in 2003. Half of the recipients served as a control group; the other half were given four shots of ALVAC-HIV, an attenuated canarypox virus carrying HIV genes, and two shots of AIDSVAX, a recombinant form of the gp120 HIV surface protein. The trial's results were published on 20 October to coincide with the AIDS vaccine meeting in Paris (S. Rerks-Ngarm *et al.* *N. Engl. J. Med.* doi:10.1056/NEJMoa0908492; 2009).

The results are a "milestone in HIV vaccine research", says first author Supachai Rerks-Ngarm, of the Thai Ministry of Public Health.

"Because the history of preventive interventions against HIV has been so poor, the HIV research community has seized on this," counters Peter Smith, a tropical epidemiologist at the London School of Hygiene & Tropical Medicine. "There is not much evidence from the data that it protects at all."

The trial was set up to measure the number of people in each group who became infected with HIV, and the amount of the virus that was circulating in the blood (the viral load) of those who became infected during the trial.



Infection rates dropped slightly during a large HIV-vaccine trial.

The teams analysed infection rates in three ways (see table). Two of those methods (known as intention to treat (ITT) and per-protocol) found a 26% drop — not statistically significant — in infection rates in the vaccine group compared with the control group. The third method (mITT), the only one presented on 24 September, excluded participants who had contracted HIV between the time they enrolled in the trial and their first vaccination. In this analysis, the reduction between the vaccine and control group infections was 31%, which just scraped into statistical significance.

This showed that statistical significance was highly dependent on whether very small numbers of individuals were excluded from either group. The small numbers of infected individuals in the trial — 132 across both groups — also meant that none of the subgroup analyses was statistically significant.

The ITT analysis is generally considered the main yardstick of the outcome of drug clinical trials, although an mITT analysis is acceptable if agreed by independent experts. For vaccine trials, the per-protocol analysis — in which patients are excluded if they don't strictly adhere

to the vaccine schedule — is the most valid, says Adel Mahmoud, former president of Merck Vaccines and now a molecular biologist at Princeton University in New Jersey.

"The results of this trial should be treated with caution and some scepticism," says Tim Peto, a researcher in tropical diseases and clinical medicine at the University of Oxford, UK. "My view is that a more balanced interpretation of the data is that the results show some evidence that the vaccine might be effective and that, unlike previous vaccine studies, this study cannot clearly rule out that the vaccine is ineffective."

Nelson Michael, a researcher at the Walter Reed Army Institute of

Research in Silver Spring, Maryland, and director of the US Military HIV Research Program, which co-organized the trial, argues that including people who didn't stick to their shots reflects a real-world vaccination scenario, and that excluding those who were HIV-positive before the trial began was justified.

The trial also failed to detect any difference between the viral load in the two cohorts. Mahmoud calls this "very, very disturbing", because an effective vaccine would be expected to at least reduce the viral load in people who become infected. However, Michael suggests that this finding might still prompt new vaccine leads, and that scientists should see whether they can uncover the immunological origin of any possible vaccine protection.

Dan Barouch, a HIV vaccine researcher at Harvard University in Cambridge, Massachusetts, says that the results are ultimately positive. "We don't understand why we saw the protection that we did, and the results are only modest, but nevertheless the Thai trial provides the first evidence of vaccine protection in humans," he says.

"Everyone is saying let's try to have hope, and this is a hope that the results mean something," says Mahmoud. "But raising expectations with no fundamental scientific base is dangerous." ■

Declan Butler

See Editorial, page 1174.

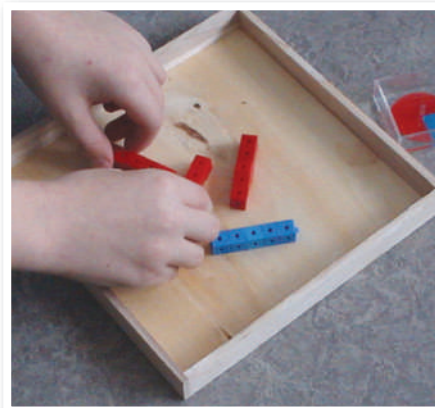
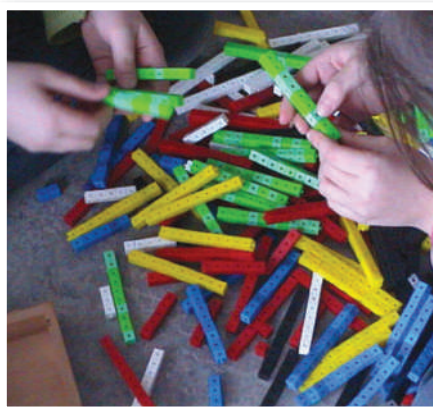
Correction

The News Feature 'Scaling the wall' (*Nature* **461**, 586–589; 2009) incorrectly located the University of Szeged in Budapest, Hungary. It is, of course, in Szeged.

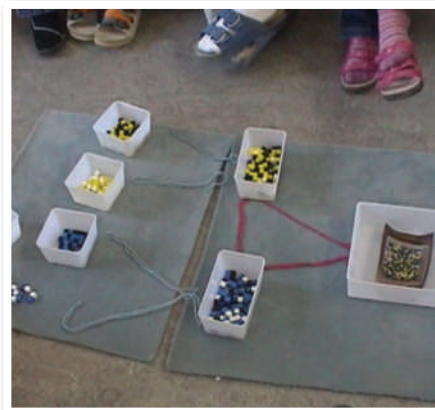
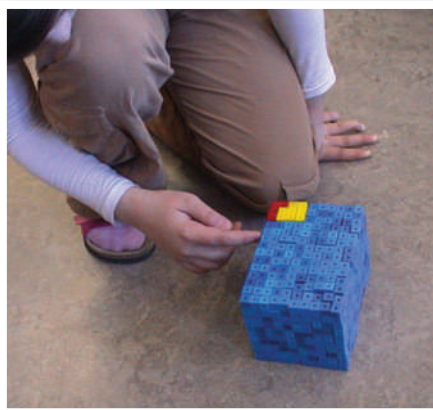
ONE TRIAL, THREE METHODS, THREE RESULTS

Method of analysis	Number of patients	Patients who became infected rate in treated vs control group	Reduction in infection
Intention to treat	16,402	56 vs 76	26.4%
Per protocol*	12,542	36 vs 50	26.2%
Modified intention to treat†	16,395	51 vs 74	31.2%

*Analysis excluded participants who failed to strictly adhere to the trial protocols, such as the calendar of vaccinations. †Analysis excluded participants who contracted HIV between the time they enrolled in the trial and their first vaccination.



RISK SCHOOL



Can the general public learn to evaluate risks accurately, or do authorities need to steer it towards correct decisions? **Michael Bond** talks to the two opposing camps.

A group of eight-year-olds sits around a classroom table, playing with coloured, plastic boxes called tinkercubes and linking them into chains. It could be playtime at almost any primary school in the world. But in this classroom, located in Stuttgart, Germany, the 'toys' are actually giving the children their first lesson in probabilistic reasoning. The cubes represent the children's attributes — red cubes for girls, blue for boys; a yellow cube attached to a red cube for a girl with glasses, a green cube attached to a blue for a boy without glasses. The students end up with a symbolic representation of their classmates as a group. And by collecting the cubes in various bins — boys versus girls, glasses versus non-glasses and so on — they begin to get a feel for the probability that, say, a boy will wear glasses or that a girl will not. It is play that is not quite play — yet the children seem hooked.

Eight might seem a little young to be learning a branch of mathematics that many students struggle to master in high school. But the idea behind the exercise — an experiment devised in 2005 by Elke Kurz-Milcke at the Institute of Mathematics and Computing in Ludwigsburg,

Germany, and tested in a number of German schools — is that earlier is better. Teaching schoolchildren how to deal with frequencies and probabilities helps to prepare them for the complexities and uncertainties of the modern world, and will help them make sound decisions throughout their lives.

That's a view strongly endorsed by Gerd Gigerenzer, a psychologist at the Max Planck Institute for Human Development in Berlin and a frequent collaborator with Kurz-Milcke. "At the beginning of the twenty-first century, nearly everyone living in an industrial society had been taught reading and writing but not how to understand information about risks and uncertainties in our technological world," he says. Earlier this year, Gigerenzer set up the Harding Center for Risk Literacy at the Max Planck Institute to try to remedy this situation. Funded for an initial five or six years by a €1.5-million (US\$2.2-million) grant from David Harding, managing director of the London-based investment-banking firm Winton

Capital and a teacher of risk communication at the University of Cambridge, UK, Gigerenzer and his team of five scientists have a twofold aim. First is to do basic research on how people perceive risk and second is to improve people's statistical and decision-making skills through education programmes.

Indeed, Gigerenzer is an outspoken advocate for the idea that people can be taught to improve their decision-making skills and has taken it upon himself to organize other researchers and set up projects. But this idea is considerably more controversial than it

"In most parts of the world, children are taught the mathematics of certainty, not of uncertainty."

— Gerd Gigerenzer

might seem. "There is a serious division in the research community," says Dan Kahan, who studies risk perception at Yale Law School in New Haven, Connecticut. He points out that many specialists in the field conclude from existing research that the public will never really be capable of making the best decision on the basis of the available scientific information. Therefore, he says, "risk decision-making should be concentrated to an even greater

E. KURZ-MILCKE

extent in politically insulated expert agencies". Those agencies, in turn, should guide or 'nudge' people into better decisions by presenting information more appropriately.

One thing both sides agree on is that poor decision-making is ubiquitous and has a serious effect on people's well-being. Faced with an unfamiliar or emotion-fraught situation, most people suspend their powers of reasoning and go with an instinctive reaction that will often lead them astray. Witness the widespread fears in the United Kingdom and the United States over the past 10 years over links between autism and the measles-mumps-rubella vaccine. Despite the lack of convincing evidence for such an association, many parents have chosen not to have their children vaccinated, leading to a rise in cases of potentially lethal measles. Likewise, a warning by the UK Committee on Safety of Medicines in 1995 that the third-generation contraceptive pill increased the risk of dangerous blood clots by 100% was followed by an additional 13,000 abortions the next year, many of them in teenage girls. The fact that the increased risk amounted to just an extra 1 in 7,000 was lost on most people — and, crucially, was not passed on by the media.

Exaggerated risk judgements also make themselves felt on environmental issues. Examples include persistent fears over the dangers of genetically modified crops in Europe, despite studies showing that the risks are considerably lower than the scare stories allege, and the hysteria triggered in the United States during the late 1980s by reports — arguably overblown and still controversial — that the plant growth regulator daminozide (Alar), used on apples and other fruit, was a potent human carcinogen. "Exaggerated risk judgements can lead to anxiety that degrades quality of life and causes excessive vigilance and self-protective behaviours," warns Ellen Peters of Decision Research, a non-profit group in Eugene, Oregon, that investigates human judgement and decision-making.

Top down

Even those who might be expected to know better — doctors, medical journalists or financial speculators, for example — often fall into the same traps as everyone else. In one experiment, Gigerenzer asked 160 gynaecologists to interpret some basic statistics about a woman's chances of having breast cancer, given that her mammography screening had come back positive. Just 21% gave the right answer¹.

"Our ability to de-bias people is quite limited," says Richard Thaler, director of the Center for Decision Research at the University of Chicago in Illinois. Thaler teaches a course in decision-making to MBA students



Genetically modified organisms (GMOs) have become the target of overexaggerated safety fears.

in their final quarter at the university's business school. Even though the students should have picked up a lot about statistics and decision-making by this time, when tested at the start of his course they exhibit all the same biases found in other groups, says Thaler. "After ten weeks of my course they do learn a bit," he says, "but I hardly turn them into rational economic decision-makers."

The problem, as many researchers in cognitive neuroscience and psychology have



Initial concerns caused many parents to deny their children the potentially life-saving measles-mumps-rubella vaccination.

concluded, is that people use two main brain systems to make decisions. One is instinctive — it operates below the level of conscious control and is often driven by emotions. The other is conscious and rational. The first system is automatic, quick and highly effective in situations such as walking along a crowded pavement, which requires the near-instantaneous integration of complex information and the carrying out of well-practised action. The second system is more useful in novel situations such as deciding on a savings plan, which calls for deliberative analysis.

Unfortunately, the first system has a way of kicking in even when deliberation would serve best. Consider a well-known example: a bat and a ball cost \$1.10 in total, the bat costs a dollar more than the ball, so how much does the ball cost? When Shane Frederick at the Massachusetts Institute of Technology in Cambridge analysed the responses to this question by nearly 3,500 individuals at eight American universities, less than half gave the right answer (5 cents)². Intuition suggests that the answer is 10 cents (it seems to fit and it feels right), and the rational system does little to correct this unless a conscious effort is made to intervene.

Such findings are why many researchers think that attempts to improve decision-making through education, which tries to put the rational system in charge of the instinctive one, lie somewhere between over-optimistic and hopeless. Two of the most prominent sceptics are Thaler and Cass Sunstein, a professor



Gerd Gigerenzer thinks that an early education in statistics will go a long way towards helping children to deal with life's uncertainties.

D. AUSSERHOFER/MPI FOR HUMAN DEVELOPMENT

at Harvard Law School who heads the White House's Office of Information and Regulatory Affairs. Thaler and Sunstein's 2008 book *Nudge* (Yale University Press) urges governments and institutions to steer people's choices in ways that should improve their lives — an approach Thaler and Sunstein call “libertarian paternalism”. Examples include automatically enrolling people into organ-donation schemes and pension plans unless they specifically choose to opt out (rather than the default being that they are not enrolled, then asking them to opt in); dollar-a-day programmes to reduce teenage pregnancies (girls receive a dollar for each day they are not pregnant); and the use of software recognition to delay the transmission of angry e-mails, giving people the option to delete before sending. In general, the idea behind the ‘nudge’ approach is to shape incentives and present information in a way that increases the chances that people will exercise good judgement.

Gigerenzer has no problem with improving the way that the information is presented. He points out that health statistics are often framed in ways that confuse not only patients but doctors, too. His Harding Center is collaborating with health insurers in Germany to persuade authorities to present health information more transparently, and he has convinced a German medical association to rewrite

one of its brochures to achieve the same kind of clarification.

But Gigerenzer is critical of those who push the nudge approach exclusively and essentially give up on people's ability to learn and reason for themselves. Some people, he says, like to attribute every poor decision to hard-wired mental processes that humans cannot control. He maintains that there is plenty of evidence that people can learn to rewire their minds — or at least, that they can learn cognitive tricks that help them to recognize and compensate for their biases. Back in the 1980s, for example, Richard Nisbett at the University of Michigan in Ann Arbor and his colleagues found that half an hour's training in statistical reasoning significantly

“Our ability to de-bias people is quite limited.”

— Richard Thaler



By giving people the chance to opt out of schemes, rather than opting in, governments can push people to make the decisions they think are right for society.

improved people's ability to rationalize everyday problems³. That included problems not generally thought of in terms of probabilities, such as whether a group's performance can be predicted from the performance of one or two of its members, or how to infer someone's personality from first impressions.

Gigerenzer's optimism about education finds cautious support from Daniel Kahneman, a senior scholar at the Woodrow Wilson School of Public and International Affairs at Princeton University in New Jersey, and a winner of the Nobel prize in economics for his pioneering work in the psychology of decision-making. “It takes an enormous amount of practice to change your intuition,” says Kahneman. “Intuition rules decision-making, that is human nature and that is how it is going to be.” Nonetheless, he says, people can improve their critical thinking so that they become better at detecting when they might make a mistake. They are then in a better position to prevent or correct it.

Instinctive bias

Researchers have found that some of the most effective cognitive tricks include looking at a problem from an outsider's perspective; considering the opposite of whatever decision you are about to make; and weighing up multiple options simultaneously rather than accepting or rejecting each one in turn⁴. Such tricks add up to what Jonathan Baron at the University of

Pennsylvania in Philadelphia calls “actively open-minded thinking” — an approach in which people intentionally look beyond the first conclusions that come to mind. He and other researchers

have found that some people are much better at this than others. “It isn't completely clear where these differences come from, but I think this kind of result is optimistic as it suggests these biases — unlike, say, [interpretation of] visual illusions — are not an unalterable part of the human condition.”

One clue to the origin of the differences comes from mathematics. Peters has found that when people with low numeracy skills are asked to assess the risks of a potential terrorist action, they are more likely than high-numeracy individuals to over-estimate the likelihood of an attack⁵. In addition, she found that numerate people are better at interpreting data about real-world scenarios, such as the performance and quality of hospitals and health insurance plans⁶.

Peters argues that people who use numbers more effectively in decision-making do so because they are better at giving numbers emotional significance and seeing them as

S. KELLY/PA WIRE

representing reality in some way — what is known as ‘affective meaning’. She suggests that it may be no coincidence that people with low numeracy skills tend to have a high body-mass index and tend to be poor at managing their own health. The challenge, says Peters, is to find a way to structure mathematics education so that students grasp the meaning in numbers faster.

This is what her colleague Paul Slovic at Decision Research calls “learning to feel the numbers”. He favours teaching children to deal with numbers in a contextual way as soon as they start to learn to count. For example, teachers should describe the number 10 in terms of something tangible — say, 10 ice-cream cones — so that children can remember the number in a way that relates to the real world. Or they could ask children to consider how it makes them feel if someone gives them a penny. What about two pence, three pence? “Get them to think about their feelings in relation to numbers and whether their feelings are logical or not,” says Slovic.



Offering teenagers a dollar for every day they are not pregnant could reduce teen pregnancy rates.

Statistical shortfall

Gigerenzer's goal is to make such ideas an integral part of education at every level. Much of his educational work is aimed at adults who deal with risk in their professional lives. The Harding Center offers training seminars in decision-making and understanding uncertainties to doctors, journalists and other specialist groups, an activity that has taken Gigerenzer around the world. His past clients include about 1,000 German gynaecologists — one-tenth of all those practising in the country — and 40 US federal judges. Of some 200 accredited law schools in the United States, he points out, only one — George Mason University School of Law in Arlington, Virginia — regularly teaches statistical thinking. “So you have an entire society, including judges and doctors, who are not being prepared for a modern technological world containing many kinds of risks,” he says.

Gigerenzer is also trying to persuade education authorities to integrate the latest findings on risk perception into school curricula, starting when children first start school and continuing right through until they leave. He is in regular contact with German education authorities, and is also working with the largest German health insurance company, AOK, to find a way to implement a programme on statistical thinking in schools in the state of Baden-Württemberg. Health insurers are interested, he says, because they realize that the health system does not run effectively,

“partly because patients don't understand the evidence”. The idea is to prepare the next generation so they know what questions to ask.

The key, he says, is for schools to teach real-world statistical problems — for example, calculating the chance that someone who tested positive for HIV actually has the virus, or comparing the dangers of riding a motorcycle in different countries. Primary schools should help pupils get used to probabilistic thinking with programmes such as Kurz-Milcke's tinker-cube exercise. “Our goal is for statistics to be taught not as a mathematical discipline, but as a problem-solving discipline,” Gigerenzer says.

Gigerenzer has had some success: several German statistics textbooks now use examples from his 2002 book *Reckoning With Risk* (Allen Lane). Furthermore, in many German states it is now compulsory to start teaching data analysis and probabilities from the first year of school. The idea is also catching on in the United States, where the National Council of Teachers of Mathematics has declared its commitment to teaching probabilities up to year 12.

Still, says Gigerenzer, there is no nationwide programme in any country that systematically teaches examples in statistics that students can usefully apply to real-life situations. And even in schools that have accepted the need for a

comprehensive education in probabilities and risks, there is often resistance from teachers who are wedded to the old system of teaching it. “In most parts of the world, children are taught the mathematics of certainty, not the mathematics of uncertainty,” he says. “Although geometry and trigonometry are beautiful systems, they are of little use in life after school compared with statistical thinking. The twenty-first century is at least as risky and uncertain as those before, and we need to prepare the next generation.”

In the end, both the education approach and the nudge approach are likely to have a role. When it comes to making better judgements, whether it's dealing with complex data or with conflicting emotional states, people — and societies — need all the help they can get. “Societally, we can do more with nudging people along, but individuals and organizations still want to think more clearly,” says Max Bazerman, who studies decision-making at Harvard Business School. With Sunstein now working within the administration of US President Barack Obama, the nudge approach seems to be gaining political capital; reforming education is proving more of a struggle.

The problem, says Gigerenzer, is as much ignorance as resistance to change among educators and policy-makers. “Often those who don't understand, don't understand that they don't understand.” But he is convinced it is worth the

fight to get the message across. He receives “a stream of letters” from mathematics teachers who have used his real-life statistical examples in their lessons and found that their students become much more interested in the subject because it applies to the world

they see around them. The long-term benefits for children could be spectacular: a statistical education that they will be able to draw on throughout their lives. The eight-year-olds puzzling over their coloured tinker-cubes in that classroom in Stuttgart should leave school well equipped to deal with the uncertainties of the modern world. ■

Michael Bond is a freelance writer based in London.

“It takes an enormous amount of practice to change your intuition.”

— Daniel Kahneman

1. Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M. & Woloshin, S. *Psychol. Sci. Publ. Int.* **8**, 53–96 (2007).
2. Frederick, S. *J. Econ. Persp.* **19**, 25–42 (2005).
3. Fong, G. T., Krantz, D. H. & Nisbett, R. E. *Cogn. Psychol.* **18**, 253–292 (1986).
4. Milkman, K. L., Chugh, D. & Bazerman, M. H. *Persp. Psychol. Sci.* **4**, 379–383 (2009).
5. Dieckmann, N. F., Slovic, P. & Peters, E. M. *Risk Anal.* **29**, 1473–1488 (2009).
6. Peters, E. et al. *J. Exp. Psychol. Appl.* **15**, 213–227 (2009).



SHOOTING PAIN

Sean Mackey inflicts pain on people in the hope of learning how to relieve it.

Erik Vance gets on the receiving end.

Outside neurology and his family, Sean Mackey doesn't have many hobbies. The one exception is his monstrous flat-screen television and large film collection. Driving to Stanford, California, on the day I am to visit Mackey's lab for testing, I am reminded of a scene from his favourite movie, *The Princess Bride*. In the film, the villain, Count Rugen, straps the hero Westley into a sinister apparatus and confesses a "deep and abiding interest in pain". Then he tortures the hero in the name of science.

It turns out that this is not far from what is in store for me.

Mackey heads the Pain Management Center at Stanford School of Medicine where, as part of his research on ways to relieve pain, he routinely inflicts it. Widely seen as one of the field's rising stars, Mackey is part of a movement to upend the way scientists look at pain, drawing the focus away from the nerves that sense it, towards the brain that processes it. His primary tool is functional magnetic resonance imaging (fMRI), which can create images of

the brain responding as the body is hurting. The trick now — and one focus of Mackey's work — is to understand whether a person can consciously change the way the brain processes and perceives pain. That's where I come in. The plan is to put me inside the fMRI scanner, apply burning heat, and see whether I can train myself to regulate my pain.

As part of his studies, Mackey has found himself struggling with a question facing the fMRI field as a whole: when is the technique ready for use outside the lab?

"I have seen pain turn people's lives upside down and absolutely destroy them."

— Sean Mackey

One of his former colleagues has started a company that plans to offer patients the fMRI 'feedback' pain-control technique that Mackey was involved in developing. But Mackey has distanced himself from the company in these

early stages, based on what he has observed elsewhere. "I've seen too many treatments that are the next latest and greatest thing out there that people get really excited about. Everybody gets on board and initially the results are fantastic. And then as time goes by we start to

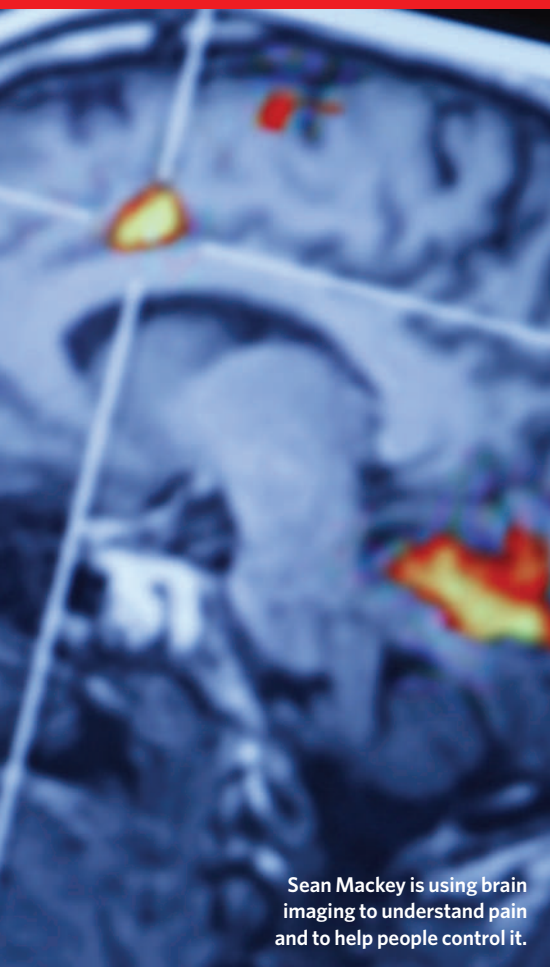
see that the results are not as good as initially proposed," he says. "And then you find out that it doesn't work at all."

Walking into his office near the Stanford Hospital, Mackey is more reminiscent of a corporate executive than a brain researcher. He is a cheery, focused ball of energy, with a quick smile and a firm, reassuring handshake. He regularly wears a suit in the lab. He says it fights the stereotype that all anaesthesiologists — he trained as one — wear "pyjamas" to work.

Destructive force

According to Mackey, the suit also tells patients that he takes his work seriously. They may be recovering from knee surgery, they may be wounded veterans or they may have a rare neurological condition that can cause excruciating full-body pain. According to the International Association for the Study of Pain, one in five people endures moderate to severe chronic pain. "I have seen it take people who are otherwise normal and turn their lives upside down and absolutely destroy them," says Mackey. Or as his colleague Ian Carroll, another Stanford anaesthesiologist, puts it: "It's

T. AVELAR/NATURE



Sean Mackey is using brain imaging to understand pain and to help people control it.

like the black hole of the brain. It dominates it and forces everything to spin around it.”

The experience of pain typically starts in receptors near the skin called nociceptors that transmit information through axon fibres to neurons in the spine, then to the brain. Until the 1990s, pain research focused mostly on nociceptors as well as neurons near the spinal cord. Pain experts would treat a backache, say, directly on the back. If they addressed the brain, it might have been with opioids, whose mechanisms were somewhat mysterious.

The arrival of the fMRI scanner changed that. The technique is an indirect measure of neural activity in the brain: as a region activates, it consumes oxygen, and neurologists use fMRI to track fresh oxygenated blood surging in to replace the old. “Imaging is now a huge part of the field,” says Allan Basbaum, editor of the journal *Pain*. Along with behavioural studies, imaging has helped to build the view that pain involves many brain areas and that chronic pain may cause long-term changes to the morphology or function of some of these regions.

As a technique, fMRI has its share of detractors. Interpreting the image relies on complicated analyses that link blood movement to a given task performed by the subject, and many neuroscience studies have come under fire for poor data analyses or interpretation. Although Mackey is a devoted user of fMRI, he is also an occasional critic. He has testified against the use of the technique in several court cases where defendants wanted to use fMRI images

to prove they were telling the truth.

“This is an interesting time in the use of this tool,” says Mackey. “You are now seeing that the application has gotten easier and easier. It’s still not quite like ordering a McDonald’s Happy Meal, where you put somebody into the scanner, press a button, and the brain pictures come out, but we are probably going to get there.”

It may seem odd to base so much of one’s scientific life on a single tool and then lobby against its application. But Mackey has a different background from many pain researchers. In addition to the MD in anaesthesiology, he has a PhD in electrical and computer engineering from the University of Arizona. He chose to work in neurology — he calls it an uncharted “Wild West” — because he could treat patients and flex his engineering muscles. This training, colleagues say, gives him a valuable understanding of the biology of pain, the complex instrumentation for measuring it and the limitations of that technology.

Personal limits

Mackey thinks that pain could be a useful proving ground for fMRI. Unlike hard-to-define cognitive or emotional states — say, deception, jealousy or anger — pain can be elicited in a controlled way at specific levels, is highly repeatable and leads to a common response: it hurts. The intensity of the pain experienced varies from person to person, but can be ranked on a scale.

The previous day, I had gone to have my own pain threshold tested. A friendly doctor took me into a small room and strapped to my arm a ‘noxious thermal stimulus,’ more accurately a ‘hot metal pad that hurts’. She added a pepper-based cream that made my skin sensitive to heat. Then she slowly worked me up to a pain level that I ranked as seven out of ten. A

seven is considered the worst pain a person can tolerate without moving, and everyone will reach their seven at a different level of heat.

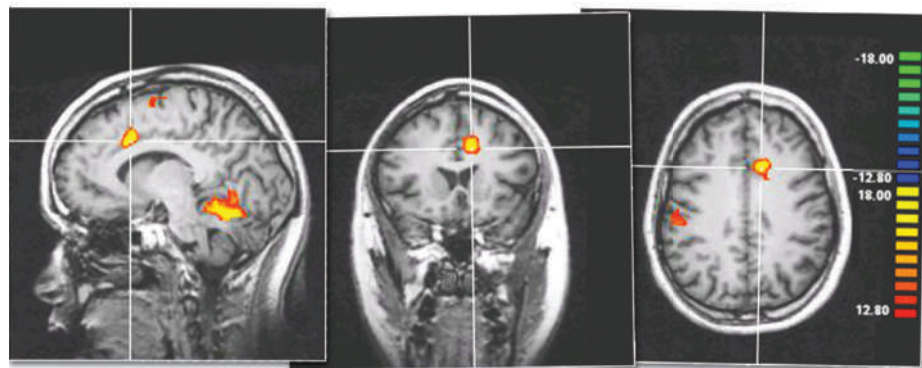
In the fMRI machine, Mackey planned to use the same burning metal plate to take me directly to seven and leave me there. I would be practising a technique that grew out of a collaboration between researchers at Stanford

and Christopher deCharms, a visiting researcher from the Massachusetts Institute of Technology (MIT) in Cambridge. The team showed people a changing image — a line graph or a picture of a fire — representing the real-time fMRI signal in their anterior

cingulate cortex, a region commonly studied in relation to pain. They showed that people could learn to manipulate the fMRI signal and their perception of pain intensity through visualization exercises, such as ‘turning down’ pain like a radio dial. The team reported in *Proceedings of the National Academy of Sciences (PNAS)* that patients with chronic pain reported on average a 64% reduction in pain on one scale¹. The study showed that fMRI could be used not only as a diagnostic, but as the means to therapy itself, and that people can exert conscious control over specific brain regions much as it is known that some people can consciously alter their heart rate.

In some of his other work, Mackey’s laboratory has used fMRI to explore these connections between pain processing and cognitive processes. Fear of pain, for example, can increase the pain itself, and Mackey’s group studied some of the brain regions involved in this anticipation². In another study³ he showed that watching someone else in pain activates brain areas that are fairly distinct from those active during one’s own pain. And in unpublished work he has found that romantic love can lessen the experience of pain. Mackey says these connections demonstrate

“There are ethical questions when you are in the middle of someone’s cognition.”
— Gary Glover



Studies suggest that people can learn to control pain using feedback to show them their brain activity.

how strong an influence conscious thought may have over pain processing.

But can this conscious control be put to use? Inside the coffin-sized tube of the fMRI machine, spasms in my back from its powerful magnet distract from the burning plate strapped again to my arm. On a screen above, I can see a squiggly line that represents the activity in a part of my anterior cingulate cortex. Mackey asks that I envision the heat as alternately searing and soothing. The aim is to master control of the line so that it (and thus my pain) goes up and down. As I switch between these visions the line on the screen twitches up and down.

It is surprisingly difficult. Willpower and meditation have little effect, and after two hours it is increasingly hard to make the stubborn little line move at all. The *PNAS* study suggests that it would probably take several sessions before I could be deft enough to reduce other pains in my body.

Commercial approach

DeCharms, first author of the *PNAS* paper, started a company called Omneuron in Menlo Park, California, to offer real-time fMRI sessions as a form of therapy. The company has already garnered a fair amount of attention and is funded by the US National Institutes of Health (NIH) in Bethesda, Maryland, to conduct a phase II clinical trial in people with pain conditions such as neuralgia, fibromyalgia or migraines. As a control, some participants will see feedback from a previous participant's brain. Omneuron is also investigating feedback fMRI to help addicts combat their cravings. DeCharms said in a short interview that feedback fMRI may someday be a valuable tool to ease chronic pain. Mackey has little to say directly about Omneuron, beyond cautiously wishing them good fortune. He adds that Stanford currently has no connections with the company's trials.

"There is something different about the balance of goals that a company has versus a nominally disinterested science project," says John Gabrieli, deCharms' former supervisor at Stanford and now at MIT. "I think that is partially why Mackey and deCharms parted ways. I think they couldn't find that balance between them."

Gary Glover, who directs Stanford's radiological-sciences lab, was a co-author on the *PNAS* paper and is more openly sceptical of attempts to commercialize feedback fMRI. He points out that the *PNAS* study determined that the participants benefited on average. This does not necessarily mean that the technique



Nature's reporter prepares for pain.

would work for any particular individual though. And customizing the feedback fMRI is generally more difficult than customizing the dose of a drug, because each person might be helped best by targeting a different brain region with a different mental exercise.

Glover also worries that with such a new technique, no one has any idea if it carries side effects: perhaps after a while the pain will worsen, or perhaps the therapy could stop patients feeling useful pain in other situations. "There are ethical questions when you are in the middle of someone's cognition," says Glover. Linda Porter, who oversees the NIH programme funding Omneuron's trial, says the obstacles are surmountable and that, compared with drug trials, the minimal risk of side effects is one of the appealing aspects of feedback fMRI.

Mackey is split on the issue of fMRI therapy. As a clinician he thinks that research should be guided by patients' needs and that new technology should be released as soon as possible. As a researcher, he stresses that it is too early

for clinical use. Pain therapies are highly subject to placebo effects, which often boost initial results and then wither away during later trials. The therapy is dependent on clinicians' ability to locate the correct part of the brain to show a patient. Much of the research so far has targeted

single regions of the brain, yet it is possible that multiple regions are involved.

Mackey admits that he doesn't know what is required to convince him fMRI therapy is ready for patients. But he is working to convince

himself. His latest study is a feedback-fMRI experiment aimed not at specific brain regions, but at the connections between them. He is also experimenting with new ways to visualize the pain and, of course, ways to apply it.

It is easy to see why some are so excited about the possibilities of fMRI therapy. It could allow the vast body of research into brain imaging to have a direct outlet that benefits patients. If it works, it would be the first non-invasive, drug-free method of directly treating specific regions of the brain, with potential applications beyond pain relief, such as addiction and depression.

But it is not without its shortcomings. After two hours in the fMRI machine, I am stiff and woozy. It is likely I have more pain than when I went in. Mackey assures me that regulating pain is in no way attached to intelligence, which seems like a polite way of saying I did not do very well. Afterwards, we sit on a shady bench and talk about the potential of the mind to regulate itself.

Still a bundle of energy, Mackey enthuses about the future of pain treatment in general, calling current technology "the dark ages". He has no doubt that self-directed treatment such as feedback fMRI is the future, perhaps done alongside drug prescriptions and physical therapy.

He adds that he is looking for volunteers for his latest round of experiments.

I stay quiet, and rub my arm a little.

Erik Vance is a freelancer writer based in Berkeley, California.

"You can't yet press a button and the brain pictures come out, but we are probably going to get there."

— Sean Mackey

1. deCharms, R. C. et al. *Proc. Natl Acad. Sci. USA* **102**, 18626–18631 (2005).
2. Ochsner, K. N. et al. *Pain* **120**, 69–77 (2006).
3. Ochsner, K. N. et al. *Soc. Cogn. Affect. Neurosci.* **3**, 144–160 (2008).

CORRESPONDENCE

Brainstem tests not adequate to diagnose death in organ donors

SIR — Your Editorial's call for serious discussion of laws governing the diagnosis of death (*Nature* **461**, 570; 2009) is most welcome, if long overdue. It is imperative that those involved in the practice of transplantation know the status of organ donors. And it is high time that those offering their organs for transplant — “after my death” are the words on NHS Organ Donor Register application forms — are clearly and fairly informed about the state they will be in if their offer is taken up.

In the United Kingdom, ‘death’ is still being certified for that purpose on the basis of purely bedside tests of some brainstem functions. There has never been sound scientific support for this standard, and it was declared “clinically dangerous” in a report by the US President's Council on Bioethics last year (see go.nature.com/58y3DP).

As you assert, few things are as sensitive as death. Its certain diagnosis is surely of such importance that it should be addressed without concern for dependent interests.

David W. Evans *Queens' College, Cambridge CB3 9ET, UK*
e-mail: dwevansmd@doctors.org.uk

Readers are welcome to join the online discussion about this Editorial at Nature Network, go.nature.com/WjUiku.

Funding on ‘Sheriff of Nottingham’ model could cut productivity

SIR — With a view to propelling Canada to the top tier of international rankings, the presidents of five large Canadian universities have proposed a radical restructuring of post-secondary education. This includes the concentration

of research in a few elite universities, with the others focusing on undergraduate education. Predictably, the country's other universities are outraged (see go.nature.com/6MMcsl).

The ‘big five’ universities (McGill, Montréal, British Columbia, Alberta and Toronto) already receive about 40% of the country's federal research funding. The ‘Sheriff of Nottingham’ model — which, like Robin Hood's adversary, robs the poor to pay the rich — relies on the premise that productivity increases faster than dollars invested. If this concept is valid, there should be more bang for the buck when funding is concentrated in a few institutions, rather than spread broadly.

Research productivity can be gauged by number of published papers and by the h-index, which measures highly cited publications, although these indicators have their flaws. I calculated both for all the researchers at 27 Canadian universities using data from the ISI Web of Science, based on papers published from 2005 to 2009. I extracted 2009 research funding per university from the three federal granting councils' websites.

According to these two indices, total research productivity and its impact both relate very strongly to the total funding received per university, but both relationships are significantly decelerating. (Details of these calculations are available from the author.) Expressed per dollar invested, research productivity and impact both decrease as funding per university increases. Bang for the research buck is better in smaller institutions.

Concentrating funding in large universities would probably increase their productivity, and perhaps their bragging rights. However, this would come at the price of reduced total research productivity summed over all Canadian universities.

The same conclusion would probably apply to any system in which finite resources for research must be divided among a pool of contenders.

David Currie *Biology Department, University of Ottawa, 30 Marie Curie Private, Ottawa, Ontario K1N 6N5, Canada*
e-mail: david.currie@uottawa.ca

A playful side to twelfth-century mathematics

SIR — As editors of the book *Lilavati's Daughters: The Women Scientists of India*, reviewed by Asha Gopinathan (*Nature* **460**, 1082; 2009), we would like to elaborate on the background to its title.

Lilavati was a mathematical treatise of the twelfth century, composed by the mathematician and astronomer Bhaskaracharya (1114–85) — also known as Bhaskara II — who was a teacher of repute and author of several other texts. The name *Lilavati*, which literally means ‘playful’, is a surprising title for an early scientific book. Some of the mathematical problems posed in the book are in verse form, and are addressed to a girl, the eponymous *Lilavati*.

However, there is little real evidence concerning *Lilavati*'s historicity. Tradition holds that she was Bhaskaracharya's daughter and that he wrote the treatise to console her after an accident that left her unable to marry. But this could be a later interpolation, as the idea was first mentioned in a Persian commentary. An alternative view has it that *Lilavati* was married at an inauspicious time and was widowed shortly afterwards.

Other sources have implied that *Lilavati* was Bhaskaracharya's wife, or even one of his students — raising the possibility that women in parts of the Indian subcontinent could have participated in higher education

as early as eight centuries ago.

However, given that Bhaskara was a poet and pedagogue, it is also possible that he chose to address his mathematical problems to a doe-eyed girl simply as a whimsical and charming literary device.

Ram Ramaswamy *Jawaharlal Nehru University, New Delhi 110 067, India*
e-mail: r.ramaswamy@mail.jnu.ac.in
Rohini Godbole *Indian Institute of Science, Bangalore 560 012, India*

Weighing up NICE against private health-care schemes

SIR — There is little question that the US health-care system requires reform. But it is debatable whether abdicating personal health-care decisions to an organization like the UK National Institute for Health and Clinical Excellence (NICE), as you recommend in your Editorials (*Nature* **461**, 315–316 and 847; 2009), is the best approach to reform.

NICE decides “which of the available medical options is most effective at treating any given condition”. This is beyond reproach and deserving of the international accolades the organization has received. However, NICE also decides “which [medical option] is worth the money”. To many Americans, this is objectionable. It is a subjective assessment intimately tied to the individual and shouldn't be in the hands of a committee.

There are myriad reform schemes being debated that still preserve an individual's control over health-care decisions. A poorly implemented private care scheme can always be reformed, but choosing government control means there is no turning back.

Todd A. Gibson *University of Colorado Denver, Box 6511, MS 8303, Aurora, Colorado 80045, USA*
e-mail: todd.gibson@ucdenver.edu

Contributions may be sent to correspondence@nature.com.

OPINION

Global Darwin: Eastern enchantment

People from Egypt to Japan used Darwin's ideas to reinvent and reignite their core philosophies and religions, says **Marwa Elshakry** in the first of four weekly pieces on how evolution was received around the world.

No other nineteenth-century scientist possessed Charles Darwin's global renown. Between the appearance of *On the Origin of Species* in 1859 and *The Descent of Man, and Selection in Relation to Sex* some 12 years later, his works were discussed in scores of languages. Darwin noted in his autobiography, published in 1887, that the theory was debated as far afield as Japan, and added with some surprise that he'd even seen an essay on the *Origin* in Hebrew showing that "the theory is contained in the Old Testament!"

His worldwide fame was, in part, thanks to technology. The first telegraphic cables were laid across the Atlantic Ocean floor around the time the *Origin* was published, and the next two decades saw Europe connected in the same way to India, China and Australasia. Meanwhile, mechanical advances in paper making and printing helped to move ideas across the globe at record speeds.

Yet the main reason for the worldwide success of Darwin's ideas was the ease with which they were assimilated into local traditions of thought — as the example of the Jewish attempt to reconcile science with scripture hints. Although Darwin himself may have found such reconciliation surprising, it was certainly not as unusual as he might have imagined. Scholars from Calcutta to Tokyo and Beijing constructed their own lineage for the theory of evolution by natural selection, tracing it to older and more familiar schools of thought and claiming ownership of what they saw as the precursors to these ideas. Although some, particularly in Europe, saw Darwin as a weapon beating down religious beliefs, around the world he was as much a force for religious resurgence and revivification as for religious scepticism. Even nineteenth-century Muslim thinkers reconciled Darwinian ideas with their own past religious and philosophical texts; which may seem ironic, given the rise of Muslim creationists today.

Cosmic order

Take as one example the work of Chinese scholar Yan Fu. In the late 1890s, Yan published a popular translation of Thomas Huxley's *Evolution and Ethics* in which he reinterpreted both Huxley and Darwin in the light of Confucian ethical debates.

Huxley, one of Darwin's most vocal



Darwin200

supporters, had argued that humans acted against the natural order of things when putting the interests of others above themselves. But for Yan, this gloomy view of nature ran counter to what he understood to be Darwin's — and Confucius's — belief

in the perfectibility of the cosmic order. Echoing older Confucian ethical debates while drawing on his own reading of Darwin and other Victorian naturalists, Yan argued that selfishness and selflessness were part of the natural order, and that each has its place in the journey towards an ideal state: the key is to achieve the right balance between the two. This was how Darwin effectively gave Yan, and many of Yan's readers, new licence to endorse one of Confucianism's ethical prescriptions.

Darwin's ideas were similarly used by late-nineteenth-century Bengali intelligentsia to support long-standing Hindu cosmological beliefs. Some of these thinkers wrote of how modern theories of positivism (the idea that true knowledge is that based on verifiable sensory experience) and evolutionism had echoes in Hindu theories of creation.

For example, Satish Mukherjee, a leading member of the Indian Positivist Society, saw Samkhya, one of the oldest schools of Hindu philosophy, as a precursor to the modern view of evolution. Under Samkhya, the world unfolds as a result of a continual cycle between creation and dissolution: consciousness, self or spirit becomes realized in matter and then separated from it, and so on. These cycles are seen to account for the creation of species as well as for the evolution of different stages of the Universe. For Mukherjee, as for many later Indian thinkers, Samkhya was therefore the theory of evolution applied to the entire cosmos.

Muslim readers found their heritage in Darwin's theory too. Supporters and critics pointed out that Muslim philosophers had long referred to the idea that species or 'kinds', as the Arabic term *anwa* suggests, could change over time. For this reason the great classics of early Muslim philosophy and cosmology were almost always cited whenever Darwin was discussed in Arabic, Farsi or Urdu.

Muslim writings from the tenth and eleventh centuries referred to a hierarchy of beings, from minerals to flora and fauna, and even argued that apes were lower forms of humans — more evidence for nineteenth-century Muslims that Darwin's theory was 'nothing new'.

Empire and evolution

One of the driving forces behind many of these scholars' work was a desire to push back against the forces of Western imperialism. At the height of European imperial power, claims about white superiority were widespread. In response, defenders of non-Western faiths drew attention to the greater rationality of their creeds to defend themselves against Western charges of backwardness and superstition. Many were keen to show that their traditions, unlike those of Western Europe, accepted, reinforced or had even anticipated the findings of modern science. By embracing Darwin's ideas, they emphasized that Christianity alone was in conflict with science.

Muhammad Abduh, the Grand Mufti of Egypt, for instance, was worried about the inroads that missionaries had made into the educational system of the Muslim Ottoman lands. He was also tired of critics pointing to Islam's supposed inability to accommodate modern pedagogy and science. In *Science and Civilization in Christianity and Islam* (1902), Abduh argued that, in contrast to Christianity,

Islam was free of the conflict with science that had so violently plagued Christian civilization in Europe. To stress this difference, he repeatedly wove references

to Darwin and evolution

into lectures on the exegesis of the Koran. Although many used Darwin to highlight the glory of their founding civilizations, they also co-opted his theory to explain their falling behind the Western world in modern times. It was seen as a way to explain both the rise of the West's technological and imperial superiority in the present, and the path to success for the rest of the world in the future.

At the height of the scramble for Africa in 1899, for instance, the Egyptian intellectual and women's-rights advocate Qasim Amin warned that "Western civilization, speeded by steam and electricity, is advancing and

"By embracing Darwin's ideas, they emphasized that Christianity alone was in conflict with science."



has expanded from its origins to all parts of the earth". The weak, he warned, would be unable to survive the onslaught. For civil servant Amin, this meant that social reform was needed. 'Self-strengthening' state reformers in Korea and Indian nationalists in the early twentieth century felt much the same way, and they too turned to evolution's advocates for instruction while pushing key governmental reforms. Of course, the battle cry of intellectuals was not always heeded.

In promoting political 'evolution', most of Darwin's proponents outside Europe subscribed not to revolution, but to change of a very gradual sort, mimicking the step-by-step slow change of natural selection.

Hiroyuki Kato, an instructor of law at the Tokyo Imperial University, used Darwin's theory to defend Japan's imperial rule at the beginning of the twentieth century. At that time, a rise of democratic movements was challenging the power of the Emperor Meiji. Kato, who also gave weekly lectures to the Emperor on constitutional and international law, supported a strongly centralized imperial line of rule. He found in Darwinism a new language in which to dress his arguments and a scientific explanation for why radical change wasn't the answer to Japan's problems.

Kato reinterpreted Darwin's 'struggle for life' as a slow, steady 'struggle for ethics'. The ethic he favoured could be counted as part of the samurai principle of self-sacrifice, which in this case he took to mean absolute allegiance to the Emperor above all other commitments. Just as through death the samurai was said to become the perfect winner, so the ultimate victor in the struggle for ethics was the

martyr dying for the sake of something bigger.

This demonstrates another characteristic common to non-European responses to Darwinism: the real question most saw lurking behind the theory of evolution was whether one could draw a moral code from nature. For Kato as for so many others, mere survival was not enough to comprise a true ethics — evolutionary or otherwise. There had to be something beyond life to give life itself a purpose. As Muslim reformer Muhammad Iqbal later put it, the main problem with Darwin's view of evolution was that it gave death 'no constructive meaning'. Perhaps for this reason, many attached their own meanings and linked Darwin to long-standing ethical systems of their own.

Paragon of scepticism?

If the ease with which Darwin's ideas were assimilated into local traditions of thought is little known today, it is because much of the discussion about Darwin in the West has focused on the supposed clash between his theory of evolution and Christianity. Certainly, ever since 1859, Darwin's name has been invoked by supporters of the forces of science in their battle against religion, and the image of Darwin as a paragon of religious scepticism has helped him to become an enduring icon of the modern sciences.

Darwin's theory did indeed help to sharpen the sense of a boundary between ideas of science and of religious faith. For disciples such as Huxley, Darwin's empirical approach offered a way to distinguish knowledge from belief, or fact from fiction. The Church of England, along with many other establishments, fought back: bishops preached that to believe

Darwin was to risk endangering one's soul.

Yet in truth, things were never this simple. Darwin was indefinite and at times inconsistent on the question of religion in his own writings. He famously left the ultimate origin of species ambiguous in the last line of the *Origin* — speaking of the power of life as 'originally breathed' into one or several forms, deploying a key Christian metaphor for creation — and he often conveyed himself as an agnostic in his letters. Not all Christians recoiled from Darwin's ideas; some Protestants and Catholics believed that they too could reconcile their doctrines with his theory and were spurred to revisit their own interpretation of scripture.

Then, as now, Darwin meant different things to different people. Globally, he was not so much a revolutionary or a scourge of faiths, as he was a revivifier of traditions. He straddled worlds between the moderns and the ancients, giving a new lease of life to ancient philosophers, ethical debates and even dynastic loyalties.

In an age in which advocates of intelligent design battle to have evolution removed from classrooms, we would do well to recall how Darwin once captured and captivated the world — not by ridding it of the forces of enchantment, faith or even God, but by revitalizing traditions of belief and re-enchanting so many.

Marwa Elshakry is associate professor of history at Columbia University, 611 Fayerweather Hall, New York, New York 10027, USA, and is the author of the forthcoming *Reading Darwin in the Middle East* (University of Chicago Press). See Editorial, page 1173. Further reading accompanies this article online. For more on Darwin see www.nature.com/darwin

ILLUSTRATION BY G. LAM

OPINION

The day the Internet age began

Forty years ago today the first message was sent between computers on the ARPANET. **Vinton G. Cerf**, who was a principal programmer on the project, reflects on how our online world was shaped by its innovative origins.

On 29 October 1969, Charley Kline, a student in the Network Measurement Center at the University of California, Los Angeles (UCLA), sent the first ever message from one computer to another on the ARPANET. The other computer was in the Stanford Research Institute, 500 kilometres to the north. Kline typed the 'l' and the 'o' of the word 'login' before one of the machines crashed. This was the inauspicious start to the ARPANET project, which led, ultimately, to the Internet.

The ARPANET was a network of dedicated telephone circuits connecting refrigerator-sized Interface Message Processors (IMPs). The IMPs formed a network that, itself, interconnected 'host' computers on which programs were implemented.

It all began in the early 1960s, with explorations of a radical communication method that came to be known as 'packet switching'. Packets are like electronic postcards with 'to' and 'from' addresses and a little text. They share communication paths the way postcards share space in a postman's van. This turns out to be more efficient for many short-lived interactions than dedicating capacity as in the telephone network with 'circuit switching'.

In the mid 1960s, Robert Taylor, director of the Information Processing Techniques Office at the US defence department's Advanced Research Projects Agency (then ARPA, and now DARPA) launched a practical packet-switching experiment largely to enable researchers at different computer-science departments to share software and resources. Led by Lawrence Roberts, the project yielded spectacular results: technologically, intellectually and philosophically.

On 2 September 1969, the first packet-switching node — or IMP — was installed at the Network Measurement Center, as the portal into the ARPANET. At the time, I was a graduate student at the centre, along with Steve Crocker and Jonathan Postel,

under director Leonard Kleinrock. Crocker led an aggregation of graduate students — called the Network Working Group — from a dozen institutions, who developed the programs used to exchange information between the computers connected by the ARPANET. Postel became the 'numbers czar' and editor of the Request for Comments series that, to this day, publishes all

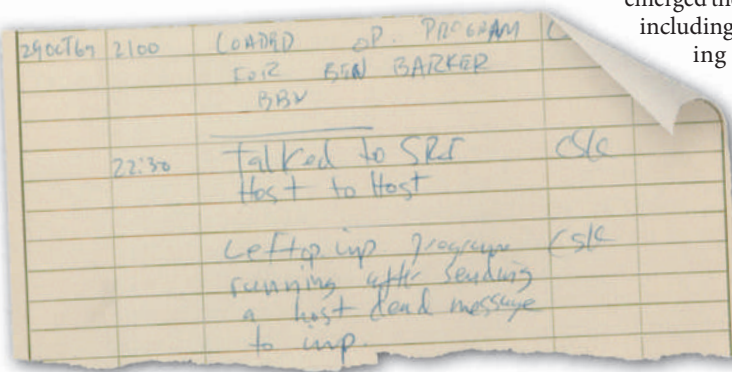
Jeff Rulifson, then a key player in the on Line System (NLS) project that was a micro-World Wide Web of its time. We were discussing the premise of computer networking. We confronted the forest of ideas and confusion of possible paths as if we were about to launch into outer space.

Out of the Network Working Group effort emerged the ARPANET host-to-host protocols including standard mechanisms for accessing computers on the other side of the network, transferring files between them as well as support for electronic mail. E-mail was an unplanned development that became extremely popular for keeping distributed researchers in close touch.

Crash testing

By December 1969 the first four IMP nodes of the ARPANET were up. The other three were at Stanford Research Institute (later SRI International) in Menlo Park, California; the University of California, Santa Barbara; and the University of Utah in Salt Lake City. Our task at UCLA was to measure performance and to compare it with Kleinrock's models. I wrote the software to inject controlled traffic into the network and to capture the associated performance data including end-to-end delays, delivery rates, effects of congestion and ability to route around failed paths in the network.

Around the turn of the year Robert Kahn and David Walden from Bolt Beranek and Newman (BBN) — the company that designed the ARPANET IMPs — came to UCLA to carry out performance tests on the network. Kahn had aired some theories about network failure modes that his colleagues thought were unlikely in practice. He and I ran a series of tests to explore these cases. Kahn would come up with a scenario and I would write the programs for the UCLA host to implement it. We managed to crash the network many times, showing where the internal control systems of the initial ARPANET implementation needed work. I remember thinking we should paint a sort of network symbol on the side of the UCLA computer for each time we crashed the network, as fighter pilots used to do with 'kills' in the Second World War.



The ARPANET logbook page recording the first computer-to-computer transmission.

Internet standards. And I became the centre's principal programmer.

Then in October came Kline's first test of what we were all working on. Happily, it was followed by many much more successful ones.

Freedom to innovate

Kleinrock focused much of his energy on mathematical modelling, and gave his graduate students remarkable freedom. We had free reign to pursue our ideas with only occasional, very strong, critical intervention by Roberts. The feeling of empowerment was reinforced by the convening of a graduate-student conference parallel to the principal investigators' meetings.

Crocker, Postel and I met scores of other students working on the project elsewhere. This led to institutional cooperation and friendships that have lasted decades. It still sends a frisson down my spine to recall tramping across an open field in Allerton House, Illinois, with Crocker and

"Preservation of the open nature of the Internet is possibly the highest imperative to emerge from this enterprise."

The extraordinary partnership forged between us has lasted for four decades and has included work on digital libraries, long-term storage and retrieval of digital content and 'knowledge robots' that can move around in the Internet performing useful functions.

Kahn organized a public demonstration of the ARPANET at the first International Conference on Computer Communication, held in Washington DC in October 1972. An IMP was installed in the ballroom of the Washington Hilton hotel, and the event went extraordinarily well. This display of the power of packet switching and computer networking, more generally, led ARPA and the Xerox Palo Alto Research Center to pursue the development of additional networks using the same principle, and spurred the growth of computer-to-computer communications.

Roberts had begun exploring the use of packet satellite communication to link network nodes via hand-held radio terminals, when terrestrial lines were unavailable or too expensive. Kahn began to consider mobile radio networks as another medium. In early 1973 the ARPANET joined two other ARPA projects: the Atlantic Packet Satellite Network (SATNET) and the Packet Radio Network (PRNET). By May 1973, Robert Metcalfe (formerly part of the Network Working Group) and David Boggs at Xerox had invented the Ethernet for local area network communications based, in part, on Metcalfe's visit to the University of Hawaii where he learned of another ARPA project, the ALOHAnet.

The Internetting problem

At around the same time Kahn described to me his vision of open network interconnection. He foresaw the need to link together packet-switched networks with different designs so that any computers could communicate freely, no matter what the communication path. Kahn started a research programme at ARPA focused on this 'Internetting' problem. Over about six months, Kahn and I met many times and by August 1973, we had developed the basic concepts of what became the transmission control protocol (TCP) and the basic architecture of the Internet. We presented a paper at a meeting at the University of Sussex in Brighton, UK, in September 1973 (published the following year as V. Cerf and R. Kahn *IEEE Trans. Commun.* **22**, 637–648; 1974) describing how to interconnect an arbitrarily large number of packet-switched networks and attached computers.

With support from ARPA, implementation of the new protocol began in 1975 at defence technology contractor BBN, Stanford and University College London. In November 1977,



For Vinton Cerf, the Internet knows no bounds.

we carried out a three-network test — a major milestone. A 'packet radio van', built by Stanford Research Institute, cruised the San Francisco Bayshore highway, radiating packets that were transferred through a gateway (a computer that links networks) to the ARPANET and relayed to University College London. There they were relayed through another gateway to the Atlantic Packet Satellite Network back across the Atlantic to the United States where they re-entered the ARPANET through another gateway and were then transported to the University of Southern California computers in Los Angeles. The distance between the packet radio van and Los Angeles was about 500 kilometres but the packets actually travelled 160,000 kilometres over two satellite hops and twice across the United States.

By 1978, these protocols, and other ARPANET application protocols (for e-mail, file transfer and remote terminal access), had been prepared to operate in a multi-network environment — the Internet. Work proceeded on implementations of these protocols for the many operating systems of machines then on the ARPANET. During 1982, a substantial effort was undertaken to implement the TCP/IP protocol suite on all the machines on all the networks.

In 1983 the ARPANET was split into a military network (MILNET) with nodes located mainly on military sites or other protected locations and the remaining ARPANET nodes located at universities, non-profit organizations, research centres and some government sites. By 1986, the US National Science Foundation had launched its NSFNET project, using the TCP/IP protocols; and other efforts were beginning to emerge overseas. The ARPANET was formally decommissioned in 1990. Hosts

were moved to sites on the NSFNET (or its regional offspring). Even the NSFNET was retired in 1995 after sufficient commercial Internet service was available to serve the academic community in the United States.

Despite the coming and going of many constituents, the Internet persists because it is an architecture for interoperability rather than a fixed set of networks.

Space next

The ARPANET project demonstrated the remarkable flexibility and utility of packet switching for computer communications. The subsequent existence of multiple networks — SATNET, PRNET, Ethernet — highlighted the versatility of the technique on different communications media. During the course of the ARPANET development, the concept of 'layering' evolved. This is the separating of various communication functions into distinguishable strata to simplify structure and implementation. It is layering that has permitted the Internet to absorb and use every new communication technology developed along the way.

Neither the ARPANET nor the Internet and its constituent networks were purpose-built for particular applications. This design philosophy has allowed these

systems to support a broad array of new techniques and applications including packetized voice, streaming video and, of course, the myriad applications built atop the World Wide

"The Internet persists because it is an architecture for interoperability rather than a fixed set of networks."

Web, under the leadership of Tim Berners-Lee, first at CERN, the European particle-physics laboratory near Geneva, now at the Massachusetts Institute of Technology in Cambridge.

Preservation of the open nature of the Internet is possibly the highest imperative to emerge from this decades-long enterprise of the research community, governments, the educational community and industry. The application space of the Internet knows no bounds. Recently, it has spawned a new round of design for an interplanetary Internet using new protocols that cope with the delay and disruption found in space communications. If you can imagine it and can program it, you can probably make it available on the Internet — a freedom to innovate that has its roots in the original ARPANET work. ■

Vinton G. Cerf is vice-president and chief Internet evangelist of Google at 1600 Amphitheatre Parkway, Mountain View, California 94043, USA. From 1969 to 1972 he was a programmer working on the ARPANET. e-mail: vint@google.com

T. BAHAR/AFP/GETTY

AUTUMN BOOKS



ILLUSTRATIONS BY JONATHAN BURTON

Reassessing the father of chemistry

Robert Boyle's character is often obscured by the shadow of Isaac Newton, but a masterful biography reveals him as larger than life, explains **Peter Anstey**.

Boyle: Between God and Science

by Michael Hunter

Yale University Press: 2009. 400 pp.
£25, \$55

In the latter half of the seventeenth century, Robert Boyle (1627–91) was the leading natural philosopher in Britain. Yet although historians have been piecing together a more-detailed profile of him in the past three decades, his popular image extends little beyond the law that bears his name and his most famous publication, *The Sceptical Chymist*. As

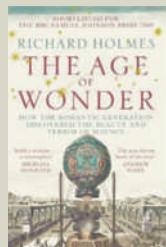
with his contemporaries Robert Hooke and Christiaan Huygens, Isaac Newton's shadow has obscured our view of Boyle. But previous biographers must share the blame for Boyle's faded image, not least the first, Thomas Birch. Writing in the 1740s with his collaborator Henry Miles, Birch removed letters and whole unpublished works from Boyle's papers in order to perpetuate the anodyne image that suited the polite tastes of the day.

Nevertheless, there is no paucity of material with which a biographer can work. Indeed, the most impressive feature of biographer Michael Hunter's *Boyle* is the meticulous care

with which he has combed the vast quantity of published and unpublished materials — including portraits, printed images and medallions — relating to Boyle's life. Hunter masterfully interweaves the narrative of Boyle's intellectual development and scientific achievements with a measured assessment of Boyle's diffident, even convoluted, personality.

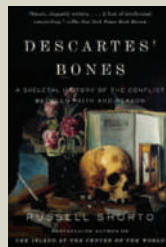
The tale begins with Boyle's domineering and ambitious father, Richard, the Earl of Cork, and moves through his infancy, childhood and Eton school years. Then follows his Grand Tour of Europe, on which Boyle had seminal experiences that were to shape his earnest Christian

NEW IN PAPERBACK



The Age of Wonder: How the Romantic Generation Discovered the Beauty and Terror of Science

by Richard Holmes (Harper Press, £9.99)
This award-winning book explores links between science and Romanticism around the start of the 1800s. "To guide readers through the science and culture of this period, Holmes masterfully dips in and out of the life of Joseph Banks," wrote David Bodanis in his review of the hardback edition (*Nature* **457**, 31–32; 2009).



Descartes' Bones: A Skeletal History of the Conflict Between Faith and Reason

by Russell Shorto (Vintage Books USA)
Russell Shorto interweaves the fate of philosopher René Descartes' bones with a narrative of Cartesian philosophy and beliefs, also exploring "the history of the uncomfortable relationship between Catholicism and Cartesianism", wrote Lisa Jardine (*Nature* **455**, 863–864; 2008).

faith and his early intellectual trajectory, and from which he emerged as a precocious adolescent.

Perhaps surprisingly, Boyle's first exploits as a writer were directed to moral and devotional topics. But at the age of 22 he was "transported and bewitched" by experimental chemistry and never looked back. So began a life dedicated to the study of nature: a life that was funded by the substantial means he inherited from his father, and that is epitomized in the title of his popular later work *The Christian Virtuoso*.

The most compelling chapters in Hunter's narrative cover Boyle's time in Oxford from the winter of 1655–56 and his emergence, in the early 1660s, as a celebrated public figure and emblem of the early Royal Society. These years were his most productive, both in terms of experimental results and written output: from 1660 to 1666, he published a dozen books at an average of 140,000 words per year. Other works took shape in this period, emerging in later decades; and still others have only recently been unearthed and published in the definitive 14-volume *The Works of Robert Boyle* (Pickering and Chatto, 1999–2000), of which Hunter is an editor.

In his publications, Boyle introduced a new and distinctive natural philosophy called corpuscularianism. He also stressed the interplay of theory and experiment in the construction of natural histories, an approach that was to dominate British science for four decades. However, most significant was the series of innovative experiments Boyle performed with his air-pump, J-tube and long pipette. Through the clever manipulation of air and mercury and with careful measurement, he established that the pressure of the air is inversely proportional to its volume. Furthermore, he solved the long-standing problem in animal physiology as to the cause of air entering the lungs in respiration: there is a differential in air pressure between the expanded lungs and the atmosphere.

Yet there is more to Boyle than the careful experimenter. Hunter shows how in the eyes of his contemporaries, from the royal court to savants abroad, Boyle was a larger-than-life character. This stemmed in part from his

overt religiosity, his reputation for professional integrity and his understated philanthropy. But it is the inner Boyle whom Hunter is most concerned to explore: Boyle the doubter, the vacillator, the stuttering and conscience-stricken man revealed in private notes written near the end of his life. Hunter displays fascination and impartiality, even wavering respect, but in the final analysis it is not clear that he really likes Boyle. However, the biographer shows maturity

by leaving the reader latitude to make up their own mind about what made Boyle tick.

This first comprehensive work on the life of Boyle is a piece of stunning scholarship, a command performance by a gifted historian. It is also a great read.

Peter Anstey teaches early modern philosophy at the University of Otago, Dunedin, New Zealand. He is the author of *The Philosophy of Robert Boyle*. e-mail: peter.anstey@otago.ac.nz

Capturing digital lives

Total Recall: How the E-Memory Revolution Will Change Everything

by Gordon Bell and Jim Gemmell
Dutton: 2009. 304 pp. \$26.95

Delete: The Virtue of Forgetting in the Digital Age

by Viktor Mayer-Schönberger
Princeton University Press: 2009.
256 pp. \$24.95

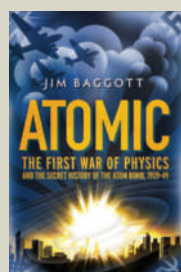
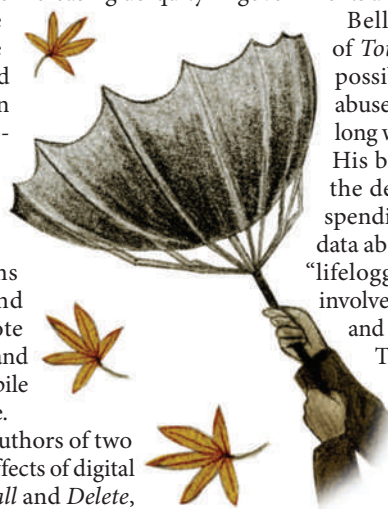
We are entering an era of unprecedented digital power. Thanks to the diminishing cost of digital storage and the increasing ubiquity of digital devices, the day is coming when we will be able to record almost every interaction we have. Digital microphones will capture our brief encounters with strangers; cameras will snap automatically as we enter new rooms or browse the web. And somewhere on a remote server farm, the images and sounds of our lives will pile up in a massive database.

On this at least, the authors of two books about the social effects of digital data storage, *Total Recall* and *Delete*,

agree. Where they differ is in what they think will happen next. Pioneering computer scientist Gordon Bell and his Microsoft colleague Jim Gemmell take a libertarian view in *Total Recall*. Digital media will free us to dip back into the past at will, they argue. Equipped with information that our brains may have lost, we will act more effectively as individuals in every part of our lives. In *Delete*, by contrast, information-policy expert Viktor Mayer-Schönberger believes people and technologies are inextricably woven into the fabric of institutions. Records of our personal data could easily make us vulnerable to the predations of governments and corporations, he warns.

Bell, the first-person narrator of *Total Recall*, acknowledges the possibility that malefactors might abuse our information, but this is a long way from his primary concern. His book is largely a chronicle of the delight he has experienced in spending the past decade gathering data about himself, a process he calls "lifelogging". At the beginning, this involved simply scanning documents and photographs onto a hard drive.

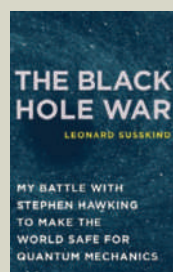
Then he began to look for ways to capture his experiences as they happened. In 2001, this resulted in the birth of the MyLifeBits project at Microsoft Research. Bell's team,



Atomic: The First War of Physics and the Secret History of the Atom Bomb, 1939–49

by Jim Baggott (Icon Books, £9.99)

Vividly written and impressively researched, *Atomic* covers the efforts of scientists and spies in the United States, Britain, the USSR and Nazi Germany to develop their own atomic weapon. Drawing on material including declassified British secret-service transcripts and documents from Soviet archives, this is a thorough but engaging account of the race to build the atomic bomb.



The Black Hole War: My Battle with Stephen Hawking to Make the World Safe for Quantum Mechanics

by Leonard Susskind (Little, Brown, £12.99)

Leonard Susskind charts his long conflict with Stephen Hawking over the fate of information in a black hole. Paul Davies's review noted it "skilfully explains the subtleties of the physics that underlie the issue, and includes anecdotes to enliven the technical details." (*Nature* 454, 579–580; 2008.)

which includes software designer and co-author Jim Gemmell, has developed a suite of digital tools for recording, storing and searching everything from old family photographs to kerbside chats.

Today, Bell wears a 'SenseCam' that automatically takes pictures every time a sensor on the device registers something that he might want recalled: a warm body, or a change in light suggesting a change of place. His desktop computer records his every keystroke. When he travels, a portable Global Positioning System continuously reports his location to MyLifeBits, which among other things allows him to log the time, date and place of any images he takes. Occasionally, he likes to play back these images in rapid-fire succession. "Talk about your life flashing before your eyes!" he enthuses.

With no shortage of hubris, Bell views the sort of tools he and his team are developing as evolutionary scaffolding. Humans are cursed with messy, organic brains that are inclined to forget, he argues. We have tried to mitigate this with memory technologies: first language, then writing, now computing. In fact, Bell asserts, "the arc of human development from the Stone Age through the present can be seen as an ongoing quest for Total Recall".

This, of course, is good news for Microsoft and the MyLifeBits team, who are developing the digital gear that makes such total recall possible. For all Bell's visionary thunder, it is hard not to hear in his writing the voice of the corporate salesman. He and Gemmell have described an imminent techno-utopia in which the technology-enabled individual moves the entire species forward to a new stage of socio-technical evolution. Yet at ground level, the book mostly offers a simple, plain-spoken

account of how we can all become better at everything if we first become better consumers of digital technology.

Mayer-Schönberger greets this vision with a polite and scholarly 'humbug'. Like Bell, he believes people will soon be able to capture almost limitless data about their lives. Unlike Bell, he does not equate memory with recall. Rather, he sees it as a process of reconstructing the past in order to act in the present. Likewise, he suggests that individual agency is more than tool-enhanced willpower. It, too,

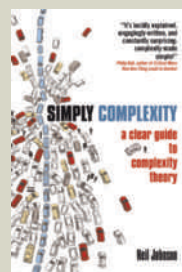
is a construction, built from both personal and social resources.

From this perspective, Mayer-Schönberger raises questions about the power of technology and how it affects our interpretation of time — a topic that Bell largely ignores. He dramatically undermines Bell's case for the value of being able to behold all the times of our lives at once. He draws on a rich body of contemporary psychological theory to argue that both individuals and societies are obliged to rewrite or eliminate elements of the past that would render action in the present impossible: a job applicant must put their former drug addiction out of their mind, a nation must stop dwelling on the cruelty of its former enemies to make peace. Here lies the central irony of our contemporary technological situation: "Through perfect memory, we may lose a fundamental human capacity — to live and act firmly in the present," he writes.

Mayer-Schönberger also fears that ubiquitous recall technologies would give institutions the power to keep a close eye on us at all times. As a result, individuals would start censoring themselves lest their actions and words be used against them, leading to a kind of social stasis. For him, Bell's proposed solution — security systems that allow users full control over their data — does not go far enough. He will only be satisfied if the devices that store the data are programmed to destroy it regularly and automatically.

If Mayer-Schönberger is right — and I'm convinced he is — then the old Kris Kristofferson song might be true after all: in the future, freedom could be just another word for nothing left to lose.

Fred Turner is assistant professor of communication at Stanford University, Stanford, California 94305-2050, USA, and author of *From Counterculture to Cyberculture*. e-mail: fturner@stanford.edu



Simply Complexity: A Clear Guide to Complexity Theory

by Neil Johnson (Oneworld, £9.99)

The science of complexity is still a fledgling field, but one that is on the rise. In this book, Neil Johnson introduces complexity, explaining what it is and how it affects us, before describing how complexity science can be used in a number of ways, from fighting disease to relationships. He also shows how, in the future, it may shed light on our understanding of quantum physics and more.



Lewis Carroll in Numberland: His Fantastical Mathematical Logical Life

by Robin Wilson (Penguin, £9.99)

Although better known for his fiction, Lewis Carroll's achievements as a mathematician should not be overlooked. Robin Wilson's book "conjures the spirit of a man who delighted in paradox yet insisted on precision ... and who wanted most of all to stump everyone he knew", wrote reviewer Jascha Hoffman (*Nature* **454**, 580–581; 2008).

Explorer of the deep

Jacques Cousteau: The Sea King

by Brad Matsen

Pantheon: 2009. 336 pp. \$27.95

Pioneer of marine conservation Jacques-Yves Cousteau — affectionately dubbed JYC, Captain Cousteau, Captain Planet, the Sea King or simply ‘the man with the red cap’ — is known worldwide for his exploration of the ocean and his success in popularizing its wonders. Yet those who mistrust his fame often question whether his work was as valuable scientifically as he made out.

Cousteau’s colourful life has already inspired several works in French and English, including two personal accounts by members of the crew of the RV *Calypso*, the legendary ship he used as an expedition vessel and research lab from 1950 until his death in 1997. The latest contribution is *The Sea King* by Brad Matsen, who has been writing books and documentary scripts about the sea and deep ocean for 30 years. Matsen pays only limited regard to Cousteau’s scientific achievements, although he claims to go further than all previous biographies of the explorer, telling the complete story of his life and throwing new light on a complex personality. In this, at least, he largely succeeds.

The book contains many little-known details of events from Cousteau’s life obtained from interviews with relatives and collaborators. Matsen accurately describes the creation of the aqualung, the first free-swimming underwater breathing equipment that Cousteau developed with the engineer Emile Gagnan in 1943, as well as his other technological contributions to underwater exploration. He describes many of the *Calypso*’s expeditions, dwelling at some length on the people who played a prominent part in them alongside Cousteau. These include Philippe Tailliez and Frédéric Dumas, known with Cousteau as the three Mousqueters, or musketeers of the sea; his first wife and business partner Simone Melchior, also

nicknamed La Bergère, or the shepherdess; his two sons; and other long-serving members of the *Calypso* crew.

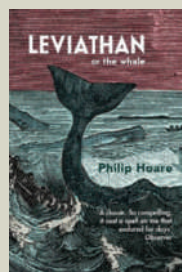
Cousteau’s story is mostly one of groundbreaking adventure, but Matsen is not afraid to delve into the dark side of his life and character, something previous biographers have been reluctant to do. There is plenty to chew on: his sometimes difficult relationships with his sons, his persistent financial problems, the fate of the *Calypso* (currently being refurbished in Brittany), the controversial role of his second wife Francine and the death of his son Philippe in a seaplane accident in 1979. He also chronicles the sad death from cancer of La Bergère, who was the soul of the *Calypso* and had a key role in many of its voyages.

One drawback of the book is that whereas Matsen covers most of the *Calypso*’s expeditions in detail, inexplicably he devotes only two lines to Cousteau’s Antarctic trip in 1972–73. According to previous accounts, the *Calypso*’s crew considered this journey their most remarkable. It was one of the last on which Cousteau was personally present, directing operations with his wife Simone. It was also the most risky: the *Calypso*’s wooden hull was not built for navigation in pack ice. The films that the crew shot of the Antarctic environment were seen by millions and the expedition was probably decisive in Cousteau’s involvement in the Protocol on Environmental Protection to the Antarctic Treaty of 1991, which designated Antarctica as a nature reserve for 50 years.



The wide popularity of Cousteau’s films and television series may have been one reason why he was sometimes accused of overplaying the scientific value of his work — a largely unfounded criticism born of jealousy and misunderstanding. True, Cousteau’s expeditions usually included professional scientists, and their contributions often came second to the storytelling. Yet he made a hugely important contribution to marine science; first, by developing technologies that enabled people to make observations and carry out experiments *in situ* underwater; second, by the interest he aroused worldwide in the sea and the damage being done to it by humans; and third, by inspiring numerous young enthusiasts to become marine scientists and professional divers.

Matsen, like so many others, is rather seduced by Cousteau’s popular image and ignores many of the scientific consequences



Leviathan: Or, the Whale

by Philip Hoare (Fourth Estate, £8.99)

Philip Hoare explores the whale and its significance to humans. Meandering through biology, economics, cultural history and his own obsession with the creatures, he describes everything from the possibility of us surviving in their bellies to gritty details about the nineteenth-century whale trade, concluding that much about the whale remains mysterious.



Witness to Extinction: How We Failed to Save the Yangtze River Dolphin

by Samuel Turvey (Oxford Univ. Press, £8.99)

Naturalist Samuel Turvey gives a personal account of the 2006 survey that determined the baiji dolphin was extinct. Describing it as a “godawful, soul-destroying experience”, he touches on the significance of the baiji’s extinction, local myths of its origin, the failed preservation project and other cetaceans such as the endangered vaquita.

of his work. He could have done more to highlight them. Most of Cousteau's first expeditions on the *Calypso* were predominantly scientific, and during the first years of his ownership she was the only French oceanographic ship, offering scientists the possibility of making direct observations down to 300 metres for the first time. Sponsored by the French National Centre of Scientific Research, his expeditions to the Mediterranean, the Red Sea and the Atlantic resulted in numerous publications,

most of which are collected in the 11 volumes of the series *Résultats Scientifiques des Campagnes de la Calypso*, which contain important contributions to marine science. Cousteau was foremost an explorer, but his contribution to science was immense. ■

Jean Vacelet is a marine biologist at the Centre d'Océanologie de Marseille, Université de la Méditerranée, Marseille, France. He took part in several of the *Calypso*'s scientific expeditions. e-mail: jean.vacelet@univmed.fr

Darwin's puppy love

Darwin's Dogs: How Darwin's Pets Helped Form a World-Changing Theory of Evolution
by Emma Townshend

Frances Lincoln: 2009. 144 pp.
\$14.95, £8.99

Even the most ardent fan of Charles Darwin might be feeling weary as his anniversary year draws to a close. Publishers have seemingly explored every corner of Darwin's life: his youth, his marriage, his attitudes to slavery and religion. Emma Townshend adds a fascinating angle — Darwin's love of dogs. Dogs were Darwin's constant companions from boyhood to old age. They were also the animals closest to hand when he explored the implications of his theories. It is surely not coincidental that Darwin's credo was "it's dogged as does it".

In *Darwin's Dogs*, Townshend adds little new to the Darwin biography. Yet her close reading of his correspondence, filtered to references to the family's dogs, produces a warmer, more intimate portrait than others so far. She plausibly claims that, aside from his years at boarding school and on the aptly named ship *HMS Beagle*, Darwin spent every day of his life in the company of dogs.

Motherless at eight years of age and packed off to boarding school, the young Darwin had, by his own admission, a "passion" for dogs,

and his letters home to his three older sisters are packed with affectionate banter about the animals. Writing of how much he missed his family's dogs, and in turn being told by his sisters how much the dogs missed him, was a face-saving way for a young man to admit his



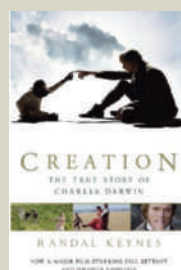
homesickness and exchange affection without embarrassment.

On his first morning back home after five years on the *Beagle*, Darwin went straight to the stables to see how his old "savage" dog, "averse to all strangers", would react to his return. Would the dog treat Darwin peaceably as befitted someone familiar, or would it growl at him showing that it had forgotten its master? As Darwin later recalled in *The Descent of Man*, the dog, "obeyed me, exactly as if I had parted with him only half an hour before. A train of old associations, dormant during five years, had thus been instantaneously awakened in his mind."

As Darwin's thoughts turned to 'transmutation' of species, the actions of dog breeders intrigued him. By carefully selecting those animals best suited to their purposes to form the parents of the next generation, breeders offered Darwin a metaphor — artificial selection — from which he could derive his great guiding principle of natural selection. Dog breeders were especially important to Darwin in trying to understand the sources of phenotypic variability and how varieties bred true to type — questions that were resolved long after Darwin's death.

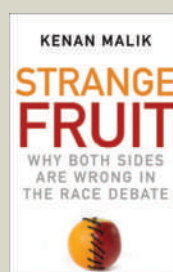
The other major issue with which Darwin grappled on his return to the United Kingdom was that of finding a wife. There too, canine thoughts were never far from his mind. When he listed for himself the pros and cons of a married life, he noted the companionship of a wife to be "better than a dog, anyhow". In Townshend's narrative, we can see in this comment affectionate praise rather than insult.

On the Origin of Species opens with a discussion of domesticated animals. When Darwin came to his magnum opus on humankind, *The Descent of Man*, and *Selection in Relation to Sex*, dogs again have centre stage. Dogs, for Darwin, know happiness and sadness, grumpiness, kindness and loyalty. They understand language — Darwin pressed his neighbour, Sir John Lubbock, into testing the latter's dog's vocabulary — and they have a sense of humour. The



Creation: The True Story of Charles Darwin
by Randal Keynes (John Murray, £7.99)

Originally titled *Annie's Box*, Randal Keynes's renamed and re-released book *Creation* focuses on Charles Darwin's relationship with his daughter Annie and how her death subsequently affected his research. "Keynes weaves a rich tapestry that gives the reader a sense of the attitudes and assumptions of the Darwin family and their class," explained Bruce Weber in a review of the hardback edition of *Annie's Box* (*Nature* 411, 739–740; 2001).



Strange Fruit: Why Both Sides Are Wrong in the Race Debate

by Kenan Malik (Oneworld, £10.99)

The subject of race is often controversial but, Kenan Malik argues, we shouldn't avoid thinking about it. He attempts to describe what race is and is not, from a biological and cultural perspective, covering modern disputes such as the US approval of a drug for African Americans with heart disease. He also looks at historical views on race and its treatment today.

concept of property ownership, Darwin argued, “is common to every dog with a bone”. With a directness and candour that still shocks today, Darwin mused that a dog’s “deep love ... for his master, associated with complete submission, some fear, and perhaps other feelings” prefigures human feelings of religious devotion.

Townshend shows a deft touch with a considerable body of Darwin scholarship. However, her simplified account of how scientific attitudes to dog behaviour have changed since Darwin is less secure. She mainly blames “behaviourists” in animal psychology for ruling inadmissible Darwin’s sympathetic attribution of human qualities to dogs, noting that anthropomorphism has been reinstated

in recent years by “cognitive psychologists”. However, the rejection of anthropomorphism was not limited to behaviourists and encompassed all forms of animal-behaviour study in the mid-twentieth century. Debate over the degree to which scientific terms used to describe human behaviour can be applied to animals continues to this day.

All in all, *Darwin’s Dogs* is thoroughly entertaining and informative. It is the ideal antidote to Darwin fatigue. ■

Clive Wynne is associate professor of psychology at the University of Florida, PO Box 112250, Gainesville, Florida 32601, USA, and author of *Do Animals Think?*
e-mail: wynne@ufl.edu

Forgotten treasure seeker

The Fossil Hunter: Dinosaurs, Evolution, and the Woman Whose Discoveries Changed The World

by Shelley Emling

Palgrave Macmillan: 2009. 256 pp.
\$27, £15.99

Remarkable Creatures

by Tracy Chevalier

Dutton/HarperCollins: 2009. 352 pp.
\$26.95/£15.99

Until recently, histories of science were written almost entirely by, for and about men. The nineteenth-century hunt for Jurassic-era fossils along the beaches of the British town of Lyme Regis was no different. Although the names of naturalists such as Georges Cuvier, William Buckland and Richard Owen who used the fossils to overturn society’s ideas about life on Earth are familiar, that of Mary Anning is only beginning to be exhumed. The publication of two books about her life — one factual, one fictional — will raise her profile in the public imagination.

Anning was a poor, working-class twelve-

year-old when she made her first major discovery within the rocks of the perilous sea cliffs in 1811: the first complete skeleton of an ichthyosaur. She went on to uncover many other important fossils, such as the first plesiosaur and the first complete skeleton of the winged reptile *Dimorphodon macronyx*. Collecting and selling small fossils to earn a living, she also led fossil hunts for naturalists visiting Lyme Regis.

Anning’s discoveries made it into the local newspapers. But it was the wealthy collectors and the established naturalists championing her finds in the halls of the Geological Society whose names became associated with them. Although her fossils helped overturn the popular idea that Earth and all its inhabitants were created in six days in 4004 BC, paving the way for Charles Darwin’s great synthesis in 1859, Anning wasn’t mentioned in key publications or lectures.

In her diligent biography *The Fossil Hunter*, Shelley Emling explains that in Anning’s day women had no place in the cut and thrust of science. Urged not to appear outdoors without a chaperone, women were barred from places where learned debates took place and were thought to lack the intellectual rigour or stamina for fieldwork. Despite this, women

did make vital contributions. Anning shared the beaches of Lyme Regis with three other female fossil collectors — the middle-class Philpott sisters, notably Elizabeth, who made well-regarded finds. And Emling describes the activity of two other talented nineteenth-century women, the wives of geologists William Buckland and Roderick Murchison, whose contributions included sketching and labelling of geological samples on expeditions.

But Anning was more than a collector or helper — she was a true scientist. She reconstructed and cleaned her own finds. She devoured scientific articles, often painstakingly copying out the entire text and figures. She engaged in spirited discussions with the men who sought her expertise and her samples. She dissected living sea creatures on the kitchen table to better understand the anatomy of their long-dead counterparts. She even conducted research, surmising, for example, that the rock-like bodies she often found within the skeletons she uncovered — coprolites — were hardened faeces. Together with William Buckland, she reconstituted coprolites in her workshop and deduced what the animals had been eating.

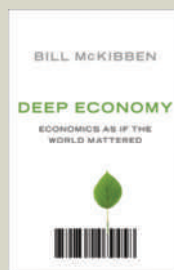
Like many retrospective narratives, Anning’s story has its heroes and anti-heroes, set-pieces and eureka moments. In *Remarkable Creatures*, Tracy Chevalier uses these devices to construct a fictionalized account of Anning’s life. Able to make things up when the details are shrouded in obscurity, Chevalier’s engaging version easily wins out over Emling’s more faithful biography. Chevalier’s unconstrained hand lets one suspend disbelief, such as in “[Buckland] asked so many questions ... that I began to feel like a pebble rolled back and forth in the tide”. By contrast, confined by the facts, Emling’s wearing reliance on the conditional



The Earth After Us: What Legacy Will Humans Leave in the Rocks?

by Jan Zalasiewicz (Oxford Univ. Press, £8.99)

Using the imagined concept of extraterrestrial beings examining Earth for evidence that humanity ever existed, geologist Jan Zalasiewicz looks at what we might leave behind in the geological record. Describing the evidence we have for Earth’s past, he explores our effects on the world and puts them in perspective over the vast timescale of the planet’s history.



Deep Economy: Economics as if the World Mattered

by Bill McKibben (Oneworld, £9.99)

Bill McKibben calls for a new focus on developing local, rather than global, economies — advocating that cities and regions should produce more of their own food, energy and culture. Such small, local economies, he argues, offer a greater sense of community and satisfaction, and better protection against an increasingly uncertain future.

perfect — in sentences such as “The bracing early-morning air would have invigorated Mary’s senses” — soon begins to grate.

Emling’s biography is the more thorough and complete work. It also frees us from the claustrophobic atmosphere of Lyme Regis, providing the context of discoveries happening elsewhere. Chevalier’s tale glosses over most of Anning’s later life. But so accurate was her fictional rendering that I felt I was reading

the same book twice. Both works did, however, lack a gripping plot. Emling’s solution was to incorporate peripheral dramas, such as natural disasters befalling Lyme Regis. Chevalier’s strategy was to sneak Elizabeth Philpott into a key session of the Geographical Society and to give Anning a putative lover.

In the end, Anning’s life story can offer no more than a pleasant but sedate read, either in fact or fiction. More evocative was the

drama, brought out well in both works, of how her discoveries shook the world: leering, nightmarish monsters materializing from the clay and hinting at a world far more ancient, savage and uncaring than anyone could possibly have imagined. ■

Jennifer Rohn is a cell biologist at University College London, UK, and editor of LabLit.com. Her first novel is *Experimental Heart*. e-mail: jenny@lablit.com

History of the hard stuff

Uncorking the Past: The Quest for Wine, Beer, and Other Alcoholic Beverages

by Patrick E. McGovern

University of California Press: 2009.

348 pp. \$29.95, £20.95

Barley, wheat and grapes in the Middle East; rice, millet and hawthorn fruit in China; figs and dates in the Levant; sorghum and palm sap in Africa; maize, cacao, cactus fruit, manioc and pepper-tree fruit in the Americas; and everywhere, honey. All these substrates were used by early humans in their quest for alcohol.

In *Uncorking the Past*, biomolecular archaeologist and University of Pennsylvania museum director Patrick McGovern argues that the desire for alcohol is innate in humans and other primates. Moreover, he believes that “the uniquely human traits” of self-consciousness, innovation, the arts and religion have been “encouraged by the consumption of an alcoholic beverage”. This is a difficult proposition to prove. In trying to do so, McGovern takes his reader on a world tour, examining the archaeological record for alcohol use across continents and cultures, searching for common themes that are indicative of universal use.

The earliest pottery artefacts with identifiable residues of a fermented beverage — arising from a mixed fermentation of rice, honey and hawthorn fruit — were found in China and date to 7000 BC. Three thousand miles away,

in the Zagros mountains in western Iran, pottery dating to 3500 BC has been found with residue of tartaric acid, indicating wine storage, as well as containers with a calcium oxalate ‘beerstone’ residue from barley beer. In Asia, Europe and the Americas, archaeologists have unearthed buildings that were constructed for the production and storage of various alcoholic beverages. And fermentation vessels and elaborate drinking sets are found in tombs of the rich and powerful across the world. Clearly, alcohol has been a part of human civilization for millennia. But has it played a part in the development of human culture?

McGovern narrates his thoughts in the first person, as if relating them to friends over a drink. He intertwines his own research findings — detailed in his earlier book, *Ancient Wine* — with those of others, and tells stories of quests to recreate ancient beverages. He describes tasting ‘Chateau Jiahu’, a modern recreation of the earliest fermentation discovered in China; and the ‘Phrygian Grog’ he named ‘Midas Touch’, a fermented beverage based on wine grapes, honey and malted barley.

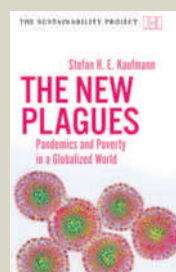
The residues of this were found in bronze containers in the burial chamber of a Phrygian king, perhaps Midas, near present-day



Blessed Days of Anaesthesia: How Anaesthetics Changed The World

by Stephanie J. Snow (Oxford Univ. Press, £9.99)

Stephanie Snow explores how early advances in anaesthetics changed society. “[This] is not a real medical history, nor is it seriously concerned with medicine or society beyond England and Scotland. But it seeks to link developments in anaesthesia with changing social, philosophical, scientific and religious attitudes in those countries,” wrote John Carmody (*Nature* **456**, 38; 2008).



The New Plagues: Pandemics and Poverty in a Globalized World

by Stefan Kaufmann (Haus Publishing, £9.99)

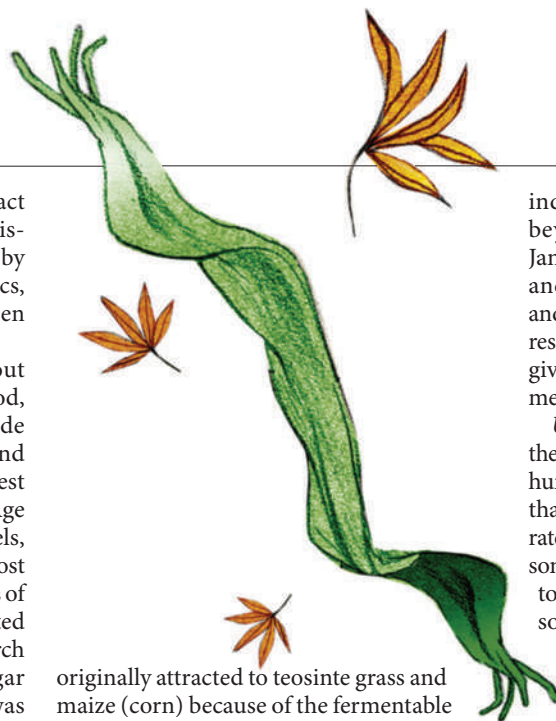
An accessible and up-to-date look at diseases that are on the rise thanks to increased globalization. Describing the various strategies that agents such as microorganisms or prions might adopt, Stefan Kaufmann delves into the conflict between rich and poor in combating outbreaks, and looks at methods for containment.

Ankara. For a general reader, the blend of fact and personal narrative is enticing, reminiscent of the mixed fermentations practised by our Neolithic ancestors; but some academics, thirsty for footnotes, may wish he had chosen a more traditional form.

McGovern begins by speculating about the role of alcohol in the Palaeolithic period, suggesting that its shamanic use, alongside other drugs, helped to develop religion and art — a proposition that is impossible to test conclusively. Neolithic and early Bronze Age cultures produced pottery and metal vessels, from which residues can be analysed. In most cases, the first fermentations were mixtures of grains, honey and wild fruit. Grains presented a problem to early brewers because the starch in the grains had to be converted to sugar before fermentation could begin. This was solved in various ways in different cultures: the use of enzymes in human saliva to break down starches is still applied in Africa and the Andes; malting and kilning of the grains is another technique, raising the possibility that beer came before bread; and the use of mould is often found in Asian rice-based brews.

Mixed fermentations, starting with higher sugar concentrations and natural yeast derived from honey and fruits, resulted in beverages of higher alcohol content than those based solely on grains. As cultures gained experience, most moved to single fermentations — beer, fruit wines or mead — with one type of beverage gaining dominance. The social importance of these beverages is reflected in the elaborate nature of fermentation vessels and drinking sets found in tombs in Asia, Europe and the Americas. Alcohol's widespread use is attested in paintings on vessels depicting communal sipping of one drink through shared straws, a scene repeated across many cultures.

The most powerful argument for alcohol as a force for innovation and social development is the claim that the initial domestication of many grains was “motivated by a desire to increase alcoholic-beverage production”, rather than to provide more food. However, only one example is explored in the book, namely the suggestion that humans were



originally attracted to teosinte grass and maize (corn) because of the fermentable sweet syrup in its stalks, and that our selection of the seed kernels of these plants for food followed only afterwards.

The broader case centres on alcohol's perceived ability to spark creativity in some

individuals, encouraging them to progress beyond tradition. In the words of William James: “Sobriety diminishes, discriminates, and says no; drunkenness expands, unites and says yes.” Although that general argument resonates with this reviewer, McGovern doesn't give any specific examples of social advancement through alcohol consumption.

Uncorking the Past doesn't prove McGovern's thesis that alcohol has been a significant force in human development, but it does demonstrate that fermented beverages have been incorporated into the fabric of society for millennia. For some, taking a ‘cup of kindness’ may be a ticket to altered consciousness; for most, it invites sociability through temporary effects to our limbic system. Both outcomes are sorely needed in today's society.

Jim Lapsley is adjunct associate professor in the Department of Viticulture and Enology at the University of California, Davis, 1 Shields Avenue, Davis, California 95616-5270, USA.
e-mail: jtlapsley@ucdavis.edu

Living by the calendar

The Seasons of Life: The Biological Rhythms that Living Things Need to Thrive and Survive

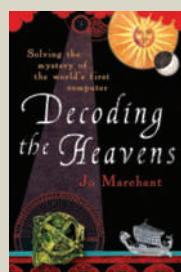
by Russell Foster and Leon Kreitzman
Profile Books/Yale University Press: 2009.
320 pp. £20/\$28

Much of biology is governed by the seasons. Reindeer seasonally adjust the colour of their eyes for better vision; newborn warblers are programmed to fly from Europe to an unknown destiny in Africa; hibernators turn down their internal thermostat for six months of the year. Most biologists would jump to unravel such seasonal feats if the time constraints were not so forbidding. Russell Foster and Leon Kreitzman lament in their latest book the slow pace of research

on annual rhythms in biology. Yet their fascinating story impresses with its wealth of facts and splendid overview.

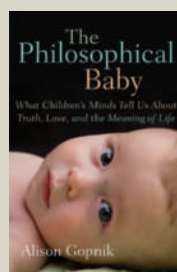
The Seasons of Life follows on from the authors' previous collaboration, *The Rhythms of Life* (Profile Books, 2004). Russell Foster, a professor in circadian neuroscience at the University of Oxford, UK, is an eminent scientist in the field of chronobiology, and a passionate one. He helped to discover specific ganglion cells in the mammalian retina that perceive light intensity and are instrumental in synchronizing biological clocks. Kreitzman is a science journalist with a lucid pen. Together, they paint a broad perspective on the functions and mechanisms of biological calendars.

The authors dedicate five chapters to the adaptation of animals and plants to the



Decoding the Heavens: Solving the Mystery of the World's First Computer

by Jo Marchant (Windmill Books, £8.99)
The 2,000-year-old Antikythera Mechanism was found in 1901, but its significance was only recently revealed. “[This] gripping and varied account will propel the mechanism to greater fame, although it may never achieve the celebrity of the Rosetta Stone that it probably deserves,” argued Andrew Robinson in his review of the hardback edition (*Nature* **455**, 867–868; 2008).



The Philosophical Baby: What Children's Minds Tell Us About Truth, Love, and the Meaning of Life

by Alison Gopnik (Bodley Head, £14.99)
Alison Gopnik's well-written book is an unusual look at the conscious mind. Drawing on research and her own pioneering studies, she reveals that processes in a baby's mind can be as complicated as those in the minds of adults, and asks what brain development can teach us about humanity.

seasons, and six to human seasonality. This preference for humans is unexpected, but appropriate and stimulating. Research interest in human seasonality has been considerable but necessarily of a descriptive nature. The annual rhythm of human reproduction, for example, is well known and has been recorded extensively in birth records. Evolutionary zoologists can only dream of having similar vast data sets.

Until recently, the consequences of birth date for human characteristics was a theme for astrologers rather than scientists. Database analyses now show that the incidence of a host of diseases later in life, such as schizophrenia, multiple sclerosis and suicidal behaviour, varies with the season of birth. This seasonal variation, the authors argue, holds the key to understanding the impact of environmental effects during gestation and early postnatal



life on adult health and lifespan.

Foster and Kreitzman mix suggestive new facts with recently recovered old references: for instance, the Babylonian king Hammurabi recommended using sunlight in the treatment of illnesses 6,000 years ago, pre-empting recent reports of the alleviation by light of depressive symptoms in seasonal affective disorder. The authors lead the reader into the literature on the systematic seasonal variations in suicide, on general mortality and on violence.

Seasonal phenomena in plants and animals are more readily approached by experimentation than are those in humans. The book tells the story of the discovery of photoperiodism — the study of the physiological changes in reaction to the length of daylight — first by Wightman Garner and Henry Allard in plants in 1920, and then by William Rowan in songbirds in 1925.

The authors also detail the

finding of innate circannual clocks: endogenous seasonal rhythms that persist even in constant temperature and day length with a usual cycle length of around 300 days rather than 365 days. Circannual clocks were first found in hibernating ground squirrels by Eric Peggelley in 1966, and in seasonally migrating songbirds in 1967 by Ebo Gwinner, the influential ornithologist to whom the book is dedicated. In the 40 years since then, significant progress has been made by only a few labs. The physiology of the

circannual pace-maker in the Soay sheep, for example, is becoming better understood through studies by Gerald Lincoln and David Hazlerigg at the Centre for Reproductive Biology in Edinburgh, UK.

Foster and Kreitzman have produced a tantalizing account of the facts behind seasonality. Its occasional nickname, 'nature's contraceptive', reflects the key function of seasonal organization: thousands of species across the globe, including those in the tropics, use seasonality to turn off reproduction at times of year when low food supply is expected and individual fitness is better served by

waiting for the next season. *The Seasons of Life* is a joy to read, and a compelling text on the importance of seasonality in the evolution of life on Earth. ■

Serge Daan holds the honorary Niko Tinbergen chair in behavioural biology at the University of Groningen, PO Box 72, 9700 AB Groningen, the Netherlands. e-mail: s.daan@rug.nl



Unmeasurable verse

Physicist and author Alan Lightman's latest work is a book-length poem. In *Song of Two Worlds*, he writes from the perspective of a man reassessing his life after a tragedy. Lightman splits his epic into two sections; in the first, he marvels at the measurable world, the glory of geometry and fact. In the second, he explores the unmeasurable, the pleasure and pain of love, the beauty of a sunset and the night sky. An excerpt from the latter section is reproduced here.

Excerpt from *Song of Two Worlds*

I am a fragment
That hurtles through space
While the breeze of the universe
Ruffles my hair.

Evening. I gaze
Through my telescope,
Searching the colors of stars.
Some are the hues of goats' wool,
Some ochre olive,
Or pink bougainvillea.

In chasms of space
I see stars born from gases,
Great thrumming furnaces oozing their heat,
Convective motions, electron opacities —
Elsewhere stars dying,
Cold cinders
Or giant explosions, eruptions of light,
Cities consumed in a nuclear blast,
Billions of years dimmed in a second.

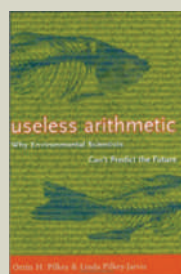
I have learned
That the heavens are violent and fragile
And doomed to destruction,
Just as this thimble the earth.
All in the cosmos is failing,
And nothing remains,
And we measure the hour of the stars,
As I measure one morning's light.

Here, in the glass of this eyepiece.

Song of Two Worlds

by Alan Lightman

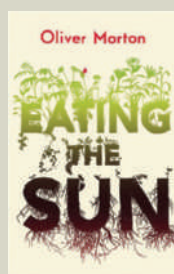
A. K. Peters: 2009. 112 pp. \$24.95



Useless Arithmetic: Why Environmental Scientists Can't Predict the Future

by Orrin H. Pilkey and Linda Pilkey-Jarvis
(Columbia Univ. Press, £15.50)

Reviewer Roger Pielke Jr wrote: "The authors have identified a critical challenge confronting the modern scientific enterprise: our ability to produce model-based predictions seems to have outpaced our ability to use such tools wisely in decision-making." (*Nature* **447**, 35–37; 2007.)



Eating The Sun

by Oliver Morton (Fourth Estate, £9.99)

There are few books on photosynthesis for the non-specialist. *Eating the Sun* fills that gap, covering the history of its discovery and its processes. "Morton's account of the ubiquitous importance of photosynthesis is an original viewpoint for looking at the world. It is written with verve and an eye for detail. His breadth of scholarship could leave other science writers green — with envy," wrote reviewer Richard Fortey (*Nature* **449**, 284–285; 2007).

NEWS & VIEWS

QUANTUM INFORMATION

Caught at the finishing line

Bob B. Buckley and David D. Awschalom

Quantum systems habitually leak information, limiting their usefulness for practical applications. By optimally reversing the leak, this information loss has been reduced to a trickle in the solid state.

The physical interactions described by quantum mechanics are fundamental to describing the world. Harnessing these quantum interactions has the potential to add a new and powerful set of tools for quantum information technology¹. Specifically, processing information using 'quantum bits' — qubits — is like adding the power of brightness and colour to each black and white (1 and 0) bit used for digital computation today. Unfortunately, just as a brightly coloured picture fades in the sun, so these quantum bits of information fade quickly through a process called decoherence². In this issue, Du and colleagues³ (page 1265) describe their work to minimize this detrimental effect in a solid-state system of electron spins (qubits) using a recently developed method that relies on repeatedly flipping the spins at well-defined intervals of time. The decohering mechanisms in the system reverse in sign when the spins are flipped, allowing most of the decoherence to simply subtract itself away.

The process that allows qubits to communicate in a useful manner also allows them to easily lose the information they encode. This information loss (decoherence) occurs when

the system interacts with itself or its surroundings through uncharacterized, random or fluctuating processes. For an ensemble of quantum objects, decoherence can manifest itself in two ways. First, the system's surroundings can vary in space, making two seemingly identical qubits perform differently through their interactions with dissimilar surroundings. And second, the surroundings can fluctuate in time, making the same qubit perform in different ways at different times. It has been shown that decoherence must be below a certain threshold for quantum computation to be feasible⁴. Thus far, it has proved difficult to physically realize a system below this threshold limit. Decoherence is therefore a central roadblock to practical quantum computation. Finding ways to understand and control this effect is vital to the future of quantum information processing.

Consider each qubit in the ensemble as a runner in a race (Fig. 1). For the purposes of quantum information, it is ideal for each runner to perform identically, with all runners crossing the finishing line simultaneously. In reality, some runners are faster than others. If one wanted them all to finish roughly at

the same time, the rules of the race could be changed so that when the race was half over, all the runners would have to turn around wherever they were and begin running back to the starting/finishing line. The more athletic members of the pack would have progressed farther and would therefore have to run just as fast as before in order to tie with their less athletic companions. This kind of reversal race can be performed on quantum systems and is called the Hahn echo⁵. It corrects for both spatial variations and variations in time that are slower than the full duration of a single race.

Of course, in reality runners do not keep a constant pace throughout a race. Some speed up, some slow down and some fluctuate in speed as the race progresses. By reversing the runners' direction multiple times during the race, it is possible to minimize the effects of their inconsistent pace, provided their speed does not fluctuate much faster than the frequency of direction flips. Analogously, with an appropriately timed set of spin-reversing pulses, Du and colleagues³ were able to reduce decoherence and extend the useful life of their qubits of choice, an ensemble of

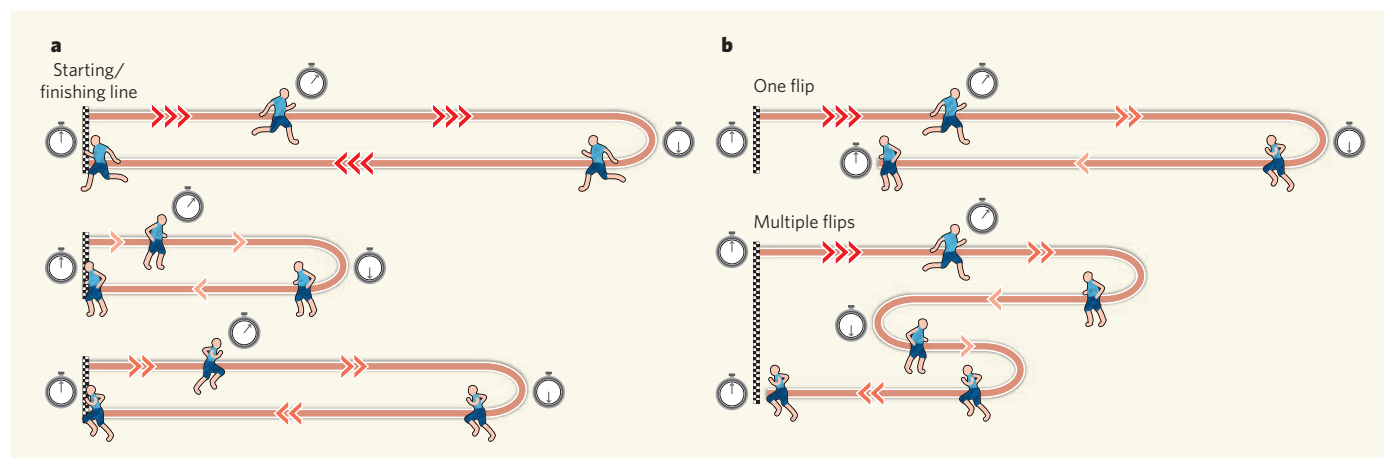


Figure 1 | A quantum race. Quantum bits in a quantum system decohere, and so lose the information they encode, much as runners in a race spread out as the race progresses. **a**, Both fast and slow runners can finish the race together by flipping everyone's running direction halfway through the race. The faster runners are consequently farther from the starting/finishing line halfway through and must run just as fast as before to tie with their slower colleagues. This approach works

provided the runners do not speed up or slow down too much during the race. **b**, By flipping the runners' direction multiple times throughout the race, inconsistent runners can be made to almost tie, provided their speed changes are less frequent than the direction flips of the race. Du and colleagues³ use an analogous optimal flipping process to repair the inconsistencies in their quantum system that lead to information loss.

electron spins in a solid-state system (a crystal of malonic acid).

The idea of using pulse sequences to reduce decoherence is not new. Besides controlling electron spins⁶ in systems such as Du and colleagues⁷, pulsing techniques have a rich history in nuclear (spin) magnetic resonance. The optimal pulse sequence studied by Du *et al.* was discovered only recently by Götz Uhrig⁷. It was then generalized^{8,9} for systems similar to that used by Du and colleagues and subsequently performed on trapped ions¹⁰. The decoherence corrections of the Uhrig pulse sequence scale linearly with the number of pulses applied. It therefore takes N pulses of flipping the spin to realize a decoherence suppression up to order N . Many of the predecessors of Uhrig's pulse sequence require an exponentially increasing number of pulses to produce the same suppression effect, requiring prohibitively large numbers of spin flips to suppress high orders of decoherence. The Uhrig pulse sequence has the ability to suppress high orders of decoherence with a feasible number of spin-flip pulses.

Du and colleagues³ explored a number of Uhrig pulse sequences in their system. Using a seven-pulse Uhrig control scheme, they were able to improve the system's coherence time by a factor of 750 compared with no control, and by a factor of 5 compared with the one-pulse Hahn-echo scheme. For up to seven pulses, they compared the Uhrig pulse scheme with other previously known pulse schemes¹¹ and showed an improvement in coherence times. By varying parameters within the sample, they were also able to isolate and study various sources of decoherence present in their system. Characterizing decoherence in similar ways in future systems could help to identify, quantify and minimize its specific forms.

Quantum control schemes such as that used by Du and colleagues should prove to be a valuable asset in understanding and attacking the decoherence of quantum information. These achievements are vital to pushing the performance of real, physical systems closer to that required for practical quantum computing. ■

Bob B. Buckley and David D. Awschalom are in the Center for Spintronics and Quantum Computation, University of California, Santa Barbara, California 93106, USA. e-mails: awsch@physics.ucsb.edu; buckley@physics.ucsb.edu

MATERIALS SCIENCE

Emerging routes to multiferroics

Ramamoorthy Ramesh

Materials that combine ferroic properties — such as ferromagnetism and ferroelectricity — are highly desirable, but rare. A new class of multiferroic solids heralds a fresh approach for making such materials.

The porous crystalline materials known as metal–organic frameworks (MOFs) are currently a hot topic of research, having found applications in catalysis¹, hydrogen storage², optical elements³ and more. Writing in the *Journal of the American Chemical Society*, Jain *et al.*⁴ describe a family of MOFs that have yet another potentially useful characteristic. They are multiferroic, combining the ferromagnetism familiar from iron bar magnets with antiferroelectricity — a property in which the molecules of a material are ordered so that adjacent molecular dipoles point in opposite directions. Multiferroics are attractive candidates for use in electrically controllable microwave elements, magnetic-field sensors and possibly even in spintronics.

Ferroic properties come in several forms, referred to as order parameters, all of which manifest themselves around some critical temperature. The primary order parameters are ferromagnetism, ferroelectricity (spontaneous electric polarization that can be reversed by an electric field) and ferroelasticity (spontaneous strain). But many other flavours also exist, including antiferroelectricity. Any material that combines more than one of these properties is described as multiferroic. If the order parameters of a multiferroic are coupled to one other, then each can be manipulated by the application of a conjugate field from the other. For example, in magnetoelectric materials, the magnetic moment of the material can be manipulated with an electric field, or the electric moment with a magnetic field. Such materials are of great fundamental scientific interest, and are also highly desirable for several applications.

Recent years have therefore seen a considerable worldwide effort to discover broad classes of multiferroics, using a combination of materials chemistry, theoretical approaches and synthetic techniques. Unfortunately, it is becoming increasingly clear that the electronic structures of molecules that are required for ferromagnetism and ferroelectricity tend to be mutually exclusive. Ferromagnetism typically requires unpaired electrons that interact through a quantum–mechanical process known as exchange coupling. But typical ferroelectric materials (such as barium titanate, BaTiO₃, and other structurally related 'perovskite' compounds that contain transition metals) require the transition-metal ion to have an empty outer shell of electrons. This fundamental contrast is the main reason why few materials are both

ferroelectric and ferromagnetic. In the materials that do have both of these order parameters, one is usually much weaker and appears at much lower temperatures than the other.

Most researchers have adopted four main approaches to try to make better magnetoelectric materials^{5–7} (Table 1). All of these attempt to introduce ferroelectricity into magnetic materials, using various mechanisms to sidestep the fundamental mismatch described above. Magnetoelectrics have also been created through a composite approach — by mixing a ferroelectric and a ferromagnet in such a way that strain is the coupling medium. Perhaps the most promising magnetoelectric material so far is bismuth ferrite (BiFeO₃, a perovskite), in which the two coupled order parameters are ferroelectricity and antiferromagnetism. Bismuth ferrite sets the benchmark in the global search for new magnetoelectric materials, guiding the design of possible multiferroic architectures, and informing the use of theoretical approaches that seek truly ferromagnetic ferroelectrics.

Jain and colleagues' MOF compounds⁴, however, represent a completely new class of multiferroics — previously reported magnetoelectric perovskites were purely inorganic compounds, but MOFs are hybrid structures comprising metal ions in complex with organic molecules. The presence of organic molecules in the structures allows hydrogen bonds to form between the MOF's components. It is these bonds that are responsible for ordering Jain and colleagues' MOFs in such a way as to engender multiferroic properties — a first in the field. The authors previously identified⁸ an antiferroelectric MOF structure that contained zinc ions (Zn²⁺). By replacing the zinc with magnetic transition-metal ions, such as iron(II) ions (Fe²⁺), they were able to make multiferroic materials.

The authors observed that, on cooling, their multiferroic MOFs undergo a transition from a paraelectric phase (in which the materials become temporarily electrically polarized in an external electric field) to an antiferroelectric phase, at critical temperatures ranging from 160 to 180 kelvin. This corresponds to a change in the molecular structure of the MOF from a disordered to a more ordered state. It is certainly nice to see that hydrogen-bonding effects can lead to relatively high transition temperatures. However, Jain *et al.* found that this transition is unaffected by a magnetic field, and so it is quite likely that there is no magnetoelectric coupling

1. Bennett, C. H. & DiVincenzo, D. P. *Nature* **404**, 247–255 (2000).
2. Zurek, W. *Phys. Today* **44** (10), 36–44 (1991).
3. Du, J. *et al.* *Nature* **461**, 1265–1268 (2009).
4. DiVincenzo, D. P. Preprint at <http://arxiv.org/abs/quant-ph/0002077v3> (2000).
5. Hahn, E. L. *Phys. Rev.* **80**, 580–594 (1950).
6. Viola, L., Knill, E. & Lloyd, S. *Phys. Rev. Lett.* **82**, 2417–2421 (1999).
7. Uhrig, G. S. *Phys. Rev. Lett.* **98**, 100504 (2007).
8. Lee, B., Witzel, W. M. & Das Sarma, S. *Phys. Rev. Lett.* **100**, 160505 (2008).
9. Yang, W. & Liu, R.-B. *Phys. Rev. Lett.* **101**, 180403 (2008).
10. Biercuk, M. J. *et al.* *Nature* **458**, 996–1000 (2009).
11. Schweiger, A. & Jeschke, G. *Principles of Pulse Electron Paramagnetic Resonance* (Oxford Univ. Press, 2001).

TABLE 1 | MECHANISMS FOR MULTIFERROICS

Mechanism	Description	Examples
Lone-pair effects	In perovskites of general formula ABX_3 , lone pairs of electrons on the A cation distort the geometry of the BX_3 anion, resulting in ferroelectricity	$BiFeO_3$, $BiMnO_3$
Geometric frustration	Long-range dipole–dipole interactions and rotations of oxygen atoms generate a stable ferroelectric state	$YMnO_3$
Charge ordering	Certain ‘non-centrosymmetric’ arrangements of ions induce ferroelectricity in magnetic materials	$LuFe_2O_4$
Magnetic ordering	Ferroelectricity is induced by magnetic long-range order in which the arrangement of magnetic dipoles lacks reflection symmetry	$TbMnO_3$, $DyMnO_3$, $TbMn_2O_4$

associated with it. But at temperatures below 10 kelvin, the compounds undergo a transition to another phase that is both antiferroelectric and weakly ferromagnetic. The fact that the transition to the magnetic state occurs at much lower temperatures is typical of ferromagnetic MOFs; this is a consequence of the indirect nature of the quantum exchange interactions between electrons that occur in these compounds.

Jain and colleagues’ compounds⁴ are a great start. But for practical applications, the transition temperatures of these multiferroic MOFs will need to be increased to around room temperature, and the strength of the coupling between the two order parameters must be increased. The beauty of MOFs is that their structures can easily be modified, which should, in principle, allow the properties of the compounds to be readily optimized. In the meantime, Jain and colleagues’ findings demonstrate the important principle that electrical order in multiferroic materials can arise from

hydrogen bonding. What’s more, in a field dominated by materials that contain the toxic element lead, MOFs open up fresh opportunities for the production of lead-free multiferroic compounds tailored for specific technological applications.

Ramamoorthy Ramesh is in the Department of Materials Science and Engineering, and the Department of Physics, University of California, Berkeley, Berkeley, California 94720, USA. e-mail: rramesh@berkeley.edu

1. Farrusseng, D., Aguado, S. & Pinel, C. *Angew. Chem. Int. Edn* **48**, 7502–7513 (2009).
2. Rowsell, J. L. C. & Yaghi, O. M. *Angew. Chem. Int. Edn* **44**, 4670–4679 (2005).
3. Allendorf, M. D., Bauer, C. A., Bhakta, R. K. & Houk, R. J. T. *Chem. Soc. Rev.* **38**, 1330–1352 (2009).
4. Jain, P. *et al.* *J. Am. Chem. Soc.* **131**, 13625–13627 (2009).
5. Ramesh, R. & Spaldin, N. A. *Nature Mater.* **6**, 21–29 (2007).
6. Cheong, S.-W. & Mostovoy, M. *Nature Mater.* **6**, 13–20 (2007).
7. Eerenstein, W., Mathur, N. D. & Scott, J. F. *Nature* **442**, 759–765 (2006).
8. Jain, P., Dalal, N. S., Toby, B. H., Kroto, H. W. & Cheetham, A. K. *J. Am. Chem. Soc.* **130**, 10450–10451 (2008).

otherwise known as the ancestor), and placed it in a simple glucose-limiting environment where it has existed ever since, being kept in a continuous state of growth by daily transfer to fresh medium⁵. At regular intervals, samples of derived populations have been collected, the reproductive success (fitness) of these types has been determined relative to the ancestor, and samples have been cryogenically stored for future reference.

Using ‘next-generation’ DNA-sequencing technologies, Barrick *et al.* determined the entire nucleotide sequence of the bacterium’s single chromosome from individual clones taken at six different time points — from one of 12 replicate populations — over the 20 years of evolution. By comparison with the ancestor, the mutations underlying 40,000 generations of evolution are revealed.

Because clones were sequenced at regular intervals, the rate of genomic evolution becomes apparent: over the first 20,000 generations, mutations accumulated at the rate of about 2 per 1,000 generations. This clock-like tempo is strongly suggestive of mode. Indeed, it is indicative of a neutral mode of evolution — evolution driven not by natural selection, but by random sampling of selectively neutral mutants⁶. According to the neutral theory⁶, mutations are expected to become standard (fixed) in populations at a constant pace. That pace is determined by the rate at which new mutations arise spontaneously in individual cells — a rate ultimately determined by the error rate of enzymes involved in DNA metabolism.

But this makes little sense, because the authors’ measures⁴ of organismal adaptation show overwhelming evidence of natural selection. In fact, during the course of the first 20,000 generations, the reproductive success of REL606 descendants improved dramatically. Moreover, the increase was strongly nonlinear. During the first 2,000 generations, fitness increased 1.5-fold relative to ancestral genotype — thereafter, the rate of improvement decreased (Fig. 1).

These discordant facts leave us in an uncomfortable position: the clock-like tempo of genomic evolution suggests a neutral mode of organismal evolution, but the tempo of organismal evolution bears the hallmarks of evolution by natural selection. Fortunately, evolution experiments with microbes offer opportunities to delve into mechanistic detail, and Barrick *et al.* do just that. They present compelling evidence that the majority of mutations arising over the first 20,000 generations of evolution are beneficial and that their fixation is due to natural selection.

In rejecting a neutral explanation for the mutations, Barrick *et al.* give life to numerous questions. The relationship between genomic and organismal (adaptive) evolution is without doubt counter-intuitive (Fig. 1). From an empirical perspective, there is a need to know whether evolution in this single replicate population

EVOLUTIONARY BIOLOGY

Arrhythmia of tempo and mode

Paul B. Rainey

An exercise in experimental evolution using bacteria has been running for more than 20 years and 40,000 generations. The results to date provide a glimpse of a new world, and are cause for both delight and unease.

In his seminal book *Tempo and Mode in Evolution*¹, palaeontologist George Gaylord Simpson argued for the value of distinguishing between the tempo of evolutionary change (the rate) and its mode (the process); moreover, he argued that tempo could be used to infer mode. Simpson’s primary interest was the large-scale variations in rate and pattern evident in the fossil record. But his thesis has been far-reaching, as influential to students of organismal evolution² as to those interested in the evolution of molecules³. Indeed, despite numerous uncertainties, molecular evolutionists use knowledge of the tempo of DNA-sequence evolution at specific loci to infer the mode of organismal evolution. On page 1243

of this issue, a group led by Richard Lenski — Barrick *et al.*⁴ — reports the results of work that provides direct insight into the rate of genomic change and organismal adaptation over 40,000 generations of evolution.

Imagine travelling back through evolutionary time, capturing at specific time intervals a complete record of the genome from a single evolving lineage, and recording all mutational events and the magnitude of their effects. Imagine that, on completion, one has a picture of not just changes in physical traits (phenotype), but also the underlying dynamic of genomic evolution. Barrick *et al.*⁴ do precisely this. In 1988, Lenski took a single archived clone of the bacterium *Escherichia coli* B (clone REL606,

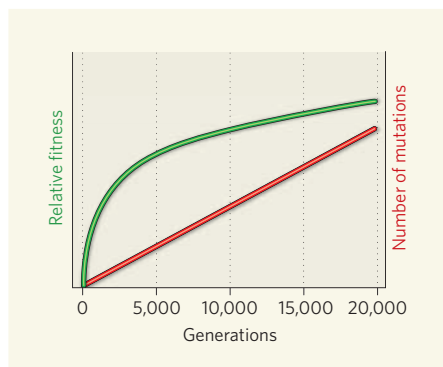


Figure 1 | Discordance in the coupling between genomic and adaptive evolution. In this summary of the tempo of genomic change (number of mutations) and organismal adaptation (fitness) over 20,000 generations of *Escherichia coli* evolution, the clock-like rate of genomic evolution revealed by Barrick and colleagues⁴ is difficult to understand given that the rate of fitness improvement decreases with time. In the absence of direct insight into genomic evolution, one would predict the rate of genomic evolution to have declined due to a reduction in the rate of appearance of new beneficial mutations, or a reduction in the average benefit of each mutation (or both).

is typical of the 11 other replicate lines. If it is typical, then one is left wondering whether an explanation will lie in the ecology of competing beneficial mutations, in subtle increases in mutation rate, or in the underlying details of the genotype-to-phenotype map. Nonetheless, the facts stand: this first glimpse into a new world will never, despite the excitement, be fully satisfactory.

The astute reader will recognize that there has thus far been no comment on the last 20,000 generations of evolution. This is intentional. After generation 20,000, a mutation arose in *mutT* — a gene encoding a protein involved in DNA repair — which caused the mutation rate to increase; the increase in mutation rate radically altered the tempo and mode of genomic evolution, although it seems to have had relatively little impact on organismal evolution. The mutations arising during the first period of evolution numbered 45 in total: over the ensuing 20,000 generations, around 600 additional mutations became fixed. The footprint left by these mutations has the signature of neutral evolution, evident primarily by the fact that many mutations had no impact on the amino-acid sequence of the encoded proteins.

The complexity of the relationship between tempo and mode of evolution at the genomic and organismal levels is the cause of some unease, and suggests that caution needs to be exercised in inferring mode of organismal evolution from rates of evolution evident in DNA. Simpson, however, would undoubtedly have relished the delight that direct knowledge of evolution on this scale brings. It is quite wonderful to see first-hand the chaos generated by the *mutT* mutation. Such insight shows just how

powerful science — and particularly the science of experimental evolution — can be, when, as Barrick *et al.*⁴ reveal, the question asked finds resonance with the technical assay.

Paul B. Rainey is at the New Zealand Institute for Advanced Study, and the Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Auckland, New Zealand.
e-mail: p.b.rainey@massey.ac.nz

1. Simpson, G. G. *Tempo and Mode in Evolution* (Columbia Univ. Press, 1944).
2. Gould, S. J. *The Structure of Evolutionary Theory* (Belknap, 2002).
3. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, 1987).
4. Barrick, J. E. *et al.* *Nature* **461**, 1243–1247 (2009).
5. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. *Am. Nat.* **138**, 1315–1341 (1991).
6. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).

ASTROPHYSICS

Most distant cosmic blast seen

Bing Zhang

The most distant γ -ray burst yet sighted is the earliest astronomical object ever observed in cosmic history. This ancient beacon offers a glimpse of the little-known cosmic dark ages.

In this issue, two papers^{1,2} report the discovery of a γ -ray burst (GRB) at a redshift of about 8.2. This is the highest redshift — or equivalently the most distant — astronomical object ever detected in the Universe. For comparison, the highest redshift recorded so far for galaxies is about 6.96 (ref. 3). For quasars — extremely bright galactic nuclei powered by supermassive black holes — the record holder is an object at 6.48 (ref. 4).

Tanvir *et al.*¹ (page 1254) and Salvaterra *et al.*² (page 1258) measured a concordant redshift for GRB 090423 on the basis of observations of its fading afterglow emission at different times after the initial burst of γ -rays. In astronomy, a larger distance (or redshift) corresponds to an earlier time in cosmic history, because it takes longer for a farther-away photon, which travels at finite speed, to reach Earth. Moreover, because the Universe is expanding, the wavelength of electromagnetic radiation emitted by a distant object is stretched to be longer and redder (redshifted) on its course to Earth. The farther away the source, the more the Universe has expanded, and therefore the higher the source's redshift. The redshift measured for GRB 090423 means that the burst occurred at a time when the Universe was about nine times smaller than it is today — putting the timing of the event at about 630 million years after the Big Bang.

GRBs are the most violent explosions in the Universe. They are believed to be associated with the formation of stellar-sized black holes or rapidly rotating, highly magnetized neutron stars during cataclysmic events such as the collapse of a massive star or the coalescence of two compact stellar objects. The reason why GRBs can outshine galaxies and quasars, which are much more massive, to become the redshift record holders (Fig. 1) is threefold.

First, thanks to their high speed (99.9995% of the speed of light), their maximum luminosity is extremely high, dwarfing that of galaxies

and quasars by many orders of magnitude.

Second, whereas galaxies and quasars look progressively dimmer at higher redshifts (both because of a larger distance from Earth and an intrinsic smaller mass at an earlier cosmic time), the apparent brightness of GRBs and their infrared afterglows (which hold the key to redshift identification) do not decrease significantly with increasing redshift. This is due to a concerted combination of two effects, the *k*-correction and the time-dilation effect⁵.

Finally, whereas bright galaxies and quasars become rarer as redshifts rise above 7, theoretical models suggest that massive stars, thought to be the progenitors of high-redshift GRBs, can form much earlier⁶. As a result, GRBs can be more easily detected at higher redshifts, and may hold the key to illuminating the cosmic 'dark ages' (Fig. 2, overleaf).

Although satellites such as Swift⁷ have been

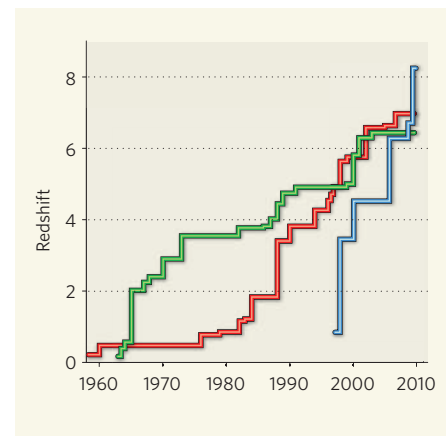


Figure 1 | Redshift ladder. Timeline of redshift record-breaking for three classes of astronomical object: galaxies (red), quasars (green) and γ -ray bursts (blue). The measurement of γ -ray-burst redshift began much later than that of galaxies and quasars, but since the first discovery it has seen a much faster pace of redshift record-breaking. (Courtesy of N. Tanvir.)

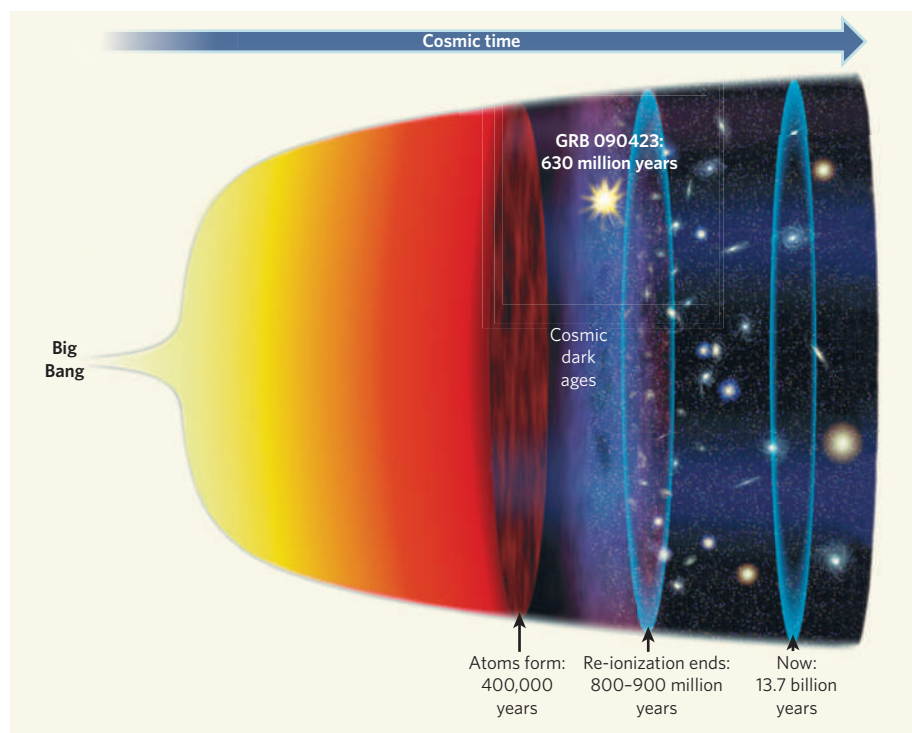


Figure 2 | The cosmic dark ages and GRB 090423. After the Big Bang, the Universe cools rapidly while expanding. About 400,000 years after this event, free electrons and protons combine to form neutral atoms, leaving a bath of background radiation that currently shines in the microwave part of the electromagnetic spectrum. Thereafter, the Universe remains neutral, until the first stars and galaxies light up at a later epoch. Photons emitted by these objects knock electrons out of atoms and 're-ionize' the Universe. Studies of the most distant galaxies and quasars suggest that the re-ionization process was completed around 800 million to 900 million years after the Big Bang, but no information is available about the cosmic 'dark ages'. Observations of γ -ray bursts such as GRB 090423 (refs 1, 2), which occurred about 630 million years after the Big Bang, offer a glimpse of the cosmic dark ages. (Adapted from ref. 15.)

successful in detecting high-redshift GRBs, characterizing these objects and measuring their redshifts is a different matter. To do so requires catching their rapidly fading afterglows and identifying their spectral signatures. The main signature is a fall in emissions, caused by intervening hydrogen-gas clouds along the light path, bluewards of the Lyman- α resonance line of hydrogen. For high-redshift GRBs, this feature is best observed with near-infrared observations, because the original ultraviolet light of such GRBs is redshifted into the observable near-infrared window.

For GRB 090423, Tanvir *et al.*¹ observed the near-infrared afterglow, starting about 17.5 hours after the burst, using the European Southern Observatory 8.2-metre Very Large Telescope in Chile. Meanwhile, Salvaterra *et al.*² observed the burst from about 14 hours after it occurred using the 3.6-metre Italian Telescopio Nazionale Galileo in Spain. Both teams discovered a clear break in emissions at wavelengths shorter than about 1.1 micrometres, which corresponds to the Lyman- α line frequency at the derived redshift.

The authors' discovery^{1,2} opens up the exciting possibility of studying the cosmic 're-ionization' epoch, and the preceding dark ages, using GRBs. In its spectrum, particularly in the absorption 'damping wing', a GRB carries information about the fraction of neutral gas

in the intergalactic medium (IGM)⁴, and so the IGM re-ionization state, at the cosmic time at which it occurred. Studying a number of GRBs spread over a range of redshifts would allow one to map out the course of the re-ionization process throughout cosmic time.

Realistically, two factors seem likely to hinder this prospect. First, afterglow observations suggest that there is a large amount of neutral gas in the interstellar medium (ISM) surrounding GRBs. In some cases, the contribution of the ISM to GRB light-absorption dominates over that of IGM absorption, making it difficult to extract information about cosmic re-ionization from the data⁸. But cosmological simulations suggest that, at high redshifts, this ISM effect should decrease with increasing redshift⁹, a trend that seems to be supported by observations of the second-highest-redshift burst, GRB 080913 (ref. 10). More high-redshift GRB data are needed to verify or disprove this prediction.

Second, afterglows fade rapidly with time. High-resolution spectroscopy is therefore needed at early (post-burst) epochs to catch them when they are still bright. In the case of GRB 090423, high-resolution spectra^{1,2} were taken from about 14 hours after the burst, when the afterglow had already faded considerably. The spectra presented by Tanvir *et al.*¹ and Salvaterra *et al.*² can therefore only serve

the purpose of redshift identification. To make further progress in studying the high-redshift Universe with GRBs, prompt alerts on high-redshift GRB candidates would be desirable. It is hoped that this could be achieved in the future by GRB space missions such as SVOM, JANUS and EXIST, or by ground-based, large near-infrared robotic telescopes.

That said, GRB 090423 still provides much information about the high-redshift Universe and the mechanisms that underlie GRB formation at that cosmic epoch. Salvaterra and colleagues² note that the burst detection may suggest that star-formation rate at high redshift is anomalously high, or that the GRB luminosity function (the relative number of GRBs that have a certain luminosity) evolves with redshift. Tanvir and colleagues¹, however, point out that its detection is consistent with the high-redshift star-formation rate predicted by some theoretical models¹¹. More data are needed to solve the authors' discrepant interpretations of the detection of GRB 090423.

Moreover, both studies^{1,2} report that the X-ray and near-infrared afterglows of the burst are not very different from those of nearer GRBs, suggesting that progenitor stars similar to those of their more recent counterparts already existed at cosmic times as early as 630 million years after the Big Bang. Detection of radio afterglow and afterglow modelling¹² suggest that the circumburst medium of GRB 090423 is also similar to that of its nearby cousins. Taken together, all of these observations^{1,2,12} indicate that the progenitor of GRB 090423 is not one of the first-generation stars, which, unlike their present-day analogues, are believed to be much more massive and metal poor (containing only hydrogen and helium)⁶.

Finally, the (rest-frame) duration of GRB 090423 is slightly longer than 1 second. Given its redshift, this is an unexpected property, but one that is shared by GRB 080913, and that implies that the burst may fall into the short-lived category. However, several arguments¹³ suggest that both GRB 090423 and GRB 080913 are related to the collapse of massive stars, rather than the merging of compact objects that is believed to power nearby, short-lived GRBs¹⁴.

The apparently short durations of these two bursts may be due to an observational-selection effect. A more intriguing possibility, which future observations could test, is that this is related to the properties of GRB progenitors at increasingly higher redshifts, which seem to produce intrinsically shorter bursts. In any case, thus far, the indication is that GRB duration alone is no longer necessarily the crucial criterion to discern the physical nature of GRBs¹³.

Bing Zhang is in the Department of Physics and Astronomy, University of Nevada, 4505 Maryland Parkway, Las Vegas, Nevada 89154-4002, USA.
e-mail: zhang@physics.unlv.edu

1. Tanvir, N. R. *et al.* *Nature* **461**, 1254–1257 (2009).
2. Salvaterra, R. *et al.* *Nature* **461**, 1258–1260 (2009).
3. Iye, M. *et al.* *Nature* **443**, 186–188 (2006).
4. Fan, X., Carilli, C. L. & Keating, B. *Annu. Rev. Astron. Astrophys.* **44**, 415–462 (2006).
5. Ciardi, B. & Loeb, A. *Astrophys. J.* **540**, 687–696 (2000).
6. Abel, T. *et al.* *Science* **295**, 93–98 (2002).
7. Gehrels, N. *et al.* *Astrophys. J.* **611**, 1005–1020 (2004).
8. Totani, T. *et al.* *Publ. Astron. Soc. Jap.* **58**, 485–498 (2006).
9. Nagamine, K., Zhang, B. & Hernquist, L. *Astrophys. J.* **686**, L57–L60 (2008).
10. Greiner, J. *et al.* *Astrophys. J.* **693**, 1610–1620 (2009).
11. Bromm, V. & Loeb, A. *Astrophys. J.* **642**, 382–388 (2006).
12. Chandra, P. *et al.* Preprint at <http://arxiv.org/abs/0910.4367> (2009).
13. Zhang, B. *et al.* *Astrophys. J.* **703**, 1696–1724 (2009).
14. Gehrels, N., Ramirez-Ruiz, E. & Fox, D. B. *Annu. Rev. Astron. Astrophys.* **47**, 567–617 (2009).
15. Bennett, J. *et al.* *Essential Cosmic Perspective* 3rd edn, 432 (Pearson Education, 2005).

CATALYSIS

Bond control in surface reactions

Jens K. Nørskov and Frank Abild-Pedersen

Catalysts steer reactions towards certain products — but the basis of their control is often unclear. Quantum chemical calculations reveal which parameters control bond formation in a network of catalytic reactions.

Control of bond formation in reactions is the key to making chemical production more energy-efficient and product-specific — thereby minimizing the formation of unwanted side products. It is also crucial for developing sustainable industrial processes for manufacturing fuels. In large-scale chemical processes, such ‘bond control’ is typically exerted by the surfaces of catalysts that consist of porous solids, or of nanoparticles supported on other materials. Writing in *Angewandte Chemie*, Loffreda *et al.*¹ report an essential step towards understanding the order in which chemical bonds form in molecules bound to a catalyst’s surface. They consider a prototype reaction in which unsaturated aldehydes react with hydrogen on a platinum surface (Fig. 1), and show that quantum chemical calculations can provide a simple set of rules to determine which bond is hydrogenated first.

Quantum chemical methods for describing surface reactions have developed extensively during the past decade, and have now reached the point at which complete catalytic reactions can be described in some detail. Indeed, the first examples in which such insight has been used to design new catalysts have been reported^{2–5}. Currently, however, only the simplest reactions can be modelled in full using quantum calculations. For complex reaction networks, a more promising approach is to develop an understanding of which factors within a network determine the overall rates. For questions relating to reaction selectivity — that is, which bond reacts first — the challenge is to identify the descriptors that determine the relative rates of all of the possible pathways in which different bonds react. These descriptors can then be calculated or measured in the search for new catalysts.

The overall reactivity of a given catalyst is traditionally understood in terms of the Sabatier principle⁶, which states that the rate of a reaction is maximized when the interactions between the intermediates and the catalyst

are neither too weak nor too strong. Weak bonding makes the catalyst too inert to allow the reaction to take place, whereas overly strong bonding may cause the catalyst to hold on to intermediates, hampering the formation

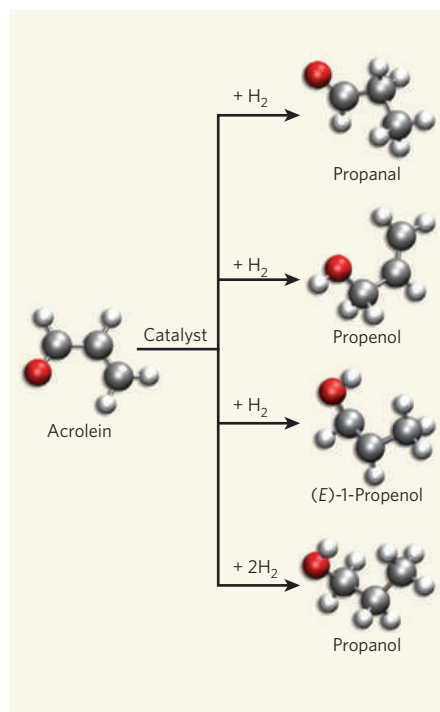


Figure 1 | Selective hydrogenation of the unsaturated aldehyde acrolein. Acrolein can, in principle, react with hydrogen in the presence of a metal catalyst to yield four possible products: propanal, a saturated aldehyde; propenol, an unsaturated alcohol; (*E*)-1-propenol, an enol; and the fully hydrogenated compound propanol, a saturated alcohol. Loffreda *et al.*¹ find that the rates of each reaction are governed entirely by the strength with which the respective intermediates bind to the surface of the catalyst. This parameter can thus be used to predict which bonds of unsaturated aldehydes will react in catalytic hydrogenations. Carbon atoms, grey; oxygen atoms, red; hydrogen atoms, white.

of products. The energies of bonds formed between the main reaction intermediates and catalytic surfaces therefore determine the overall catalytic rate, and can be used as descriptors of that rate⁷. Only a few bond energies are usually needed to determine a catalyst’s reactivity, because the various parameters associated with reactivity tend to scale with each other, effectively limiting the number of independent variables^{2,7}. Specifically, the activation energies of reactions (the energies required to initiate reactions) generally scale with the energy difference between reaction intermediates, and surface-bond energies for different intermediates in a reaction pathway often scale with each other.

Loffreda *et al.*¹ report that specific surface-bond energies also work as descriptors for reaction selectivity. They first show that the order in which the different double bonds of acrolein (a simple unsaturated aldehyde, C₃H₄O) are hydrogenated on the surface of a metal such as platinum can be understood by dividing all of the possible reactions that could occur into four groups. Each group is defined by its reaction centre — one of the three carbon atoms, or the oxygen atom. The authors find that, for each group and at any given point along a reaction pathway, the transition-state energy (the energy at which the complex of reactants becomes more product-like than reactant-like) scales linearly with the surface-bond energy of the preceding intermediate–metal complex. In fact, there is an almost one-to-one correlation between the energies. This means that the activation energies for all of the different reactions that can take place at each centre are about the same, something that has also been found in studies of smaller molecules reacting on a variety of transition metals⁸.

Intriguingly, Loffreda *et al.*¹ go on to show that the same scaling relationship also applies to another unsaturated aldehyde, prenal. This means that the transition-state energies for prenal hydrogenation could have been predicted by calculating the surface-bond energies for the relevant reaction centres of that molecule. For any given reaction that has several possible reaction pathways, the one that has the lowest transition-state energy typically has the fastest rate — it will occur before the others. The authors have therefore shown that surface-bond energy is a descriptor of reaction selectivity. If this is a general result, then it will allow the reaction selectivities for hydrogenations of many aldehydes on many different catalysts to be screened rapidly using quantum chemical calculations. Hydrogenations are exceptionally useful reactions in industry, so this is an exciting prospect.

Loffreda and colleagues’ work exemplifies the great progress that has been made in our understanding of solid-state catalysts in recent years as a direct result of quantum chemical calculations. Extending the use of computational techniques to complex reaction networks is the next great challenge — but there are already

promising signs that this can be done, and that these techniques will become an essential tool for catalyst design. To fulfil this promise, improvements to the accuracy of quantum chemical calculations are needed, as well as faster computers and algorithms to enable calculations for systems of greater complexity. But perhaps more critically, as Loffreda *et al.*¹ show, we need the insight provided by computational techniques to distinguish between the important and the less important parameters of catalysis.

Jens K. Nørskov and Frank Abild-Pedersen are in the Department of Physics, Center for Atomic-

scale Materials Design, Technical University of Denmark, 2800 Kongens Lyngby, Denmark.
e-mails: nørskov@fysik.dtu.dk;
abild@fysik.dtu.dk

1. Loffreda, D., Delbecq, F., Vigné, F. & Sautet, P. *Angew. Chem. Int. Edn* doi:10.1002/anie.200902800 (2009).
2. Nørskov, J. K., Bligaard, T., Rossmeisl, J. & Christensen, C. H. *Nature Chem.* **1**, 37–46 (2009).
3. Linic, S., Jankowiak, J. & Barteau, M. A. *J. Catal.* **224**, 489–493 (2004).
4. Toulhoat, H. & Raybaud, P. *J. Catal.* **216**, 63–72 (2003).
5. Greeley, J. & Mavrikakis, M. *Nature Mater.* **3**, 810–815 (2004).
6. Balandin, A. A. *Adv. Catal.* **19**, 1–210 (1969).
7. Abild-Pedersen, F. *et al. Phys. Rev. Lett.* **99**, 016105 (2007).
8. Liu, Z.-P. & Hu, P. *J. Am. Chem. Soc.* **125**, 1958–1967 (2003).

STRUCTURAL BIOLOGY

DNA binding shapes up

Tom Tullius

DNA-binding proteins have the daunting task of finding their binding sites among the 3 billion base pairs of the human genome. The shape of DNA, and not just its sequence, may offer proteins much-needed direction.

The genetic information embodied in DNA must be decoded at the right time and in the right type of cell. To achieve this, proteins that control such processes have to bind to specific places in the genome. How a protein finds the correct spot to bind to among all the possible sites (3 billion base pairs in the human genome, for example) has been the preoccupation of molecular and structural biologists for decades.

Watson and Crick taught us that DNA adopts the form of a double helix, and much of DNA's biological function is evident from the complementary strands that make up this iconic helix. These days, we tend to think of DNA as a string of letters (A, G, C and T), which stand for the bases of the four nucleotides that make up the DNA polymer. The familiar genetic code consists of sets of three DNA nucleotides that specify one of the 20 amino acids that make up proteins; the deciphering of this code was one of the triumphs of the early days of molecular biology.

One might imagine that a protein could recognize its binding site in the genome by somehow 'reading' these letters, and in fact this has been found to be the case from studying structures of protein–DNA complexes. The DNA double helix has two grooves, a major and a minor one, that wind around the central axis of the molecule, and reading is achieved using hydrogen bonds that form between protein side chains and the edges of the DNA nucleotides that are exposed in the major groove. But unlike the genetic code, a simple code for protein recognition of DNA has not emerged despite years of effort.

Perhaps we lose something in the simplification of DNA to a one-dimensional string

of letters. We may forget that DNA is a molecule with a three-dimensional shape that is not perfectly uniform. Rohs *et al.*¹ (page 1248 of this issue) now find that one structural feature of DNA, the shape of its minor groove, can be exploited by proteins for specific recognition.

A particular sequence of nucleotide letters presents a unique array of hydrogen-bond donors and acceptors in the major groove, providing a clear mechanism for reading that sequence². The minor groove, though, has been inscrutable, as a simple code of hydrogen-bond donors and acceptors that specify a nucleotide sequence is not present in this groove.

The minor groove has another trick up its sleeve — its width varies depending on which nucleotides are present in a segment of DNA. And the width of the minor groove has a physical consequence that goes beyond the merely structural, stemming from the charged groups (phosphates) that are arrayed along the outer edge of the DNA backbone, one per nucleotide (Fig. 1a). Where the minor groove is narrow, the electric-field lines due to the negatively charged phosphates are focused into the groove, leading to an enhanced negative electrostatic potential in that segment of the double helix.

Rohs and colleagues¹ exhaustively searched the databases of three-dimensional structures of protein–DNA complexes and found many examples of proteins that use amino acids containing positively charged side chains, principally arginines, to read the electrostatic potential of the minor groove. Where the groove width and electrostatic potential are optimal, an arginine side chain of a DNA-binding protein is often seen to sit snugly in the minor groove (Fig. 1b). It is important to appreciate

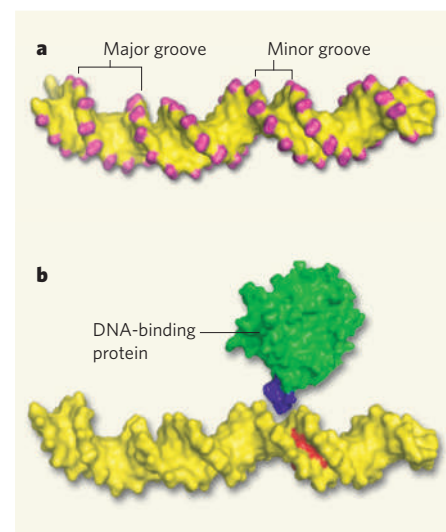


Figure 1 | Getting into the groove. Rohs *et al.*¹ report that the shape of the minor groove of DNA can direct the binding of proteins to specific sites. **a**, Negatively charged phosphate groups (magenta) are arrayed along the outer edge of the DNA major and minor grooves that spiral around the axis of the double helix. The width of the minor groove varies depending on the sequence of nucleotides. This variation leads to differences in the distance between phosphates across the groove, which in turn lead to variation in the negative electrostatic potential along the minor groove. **b**, A representation of a DNA-binding protein (green) that has a positively charged side chain on its surface, for example arginine (purple), is shown. The protein may recognize a binding site on DNA by its electrostatic potential. The protein is about to bind to the segment of the DNA minor groove that has the optimum groove width and electrostatic potential for binding (red). The DNA structure in **a** and **b** is derived from a structure in the RCSB Protein Data Bank, accession number 2O61. The illustration of a DNA-binding protein in **b** is hypothetical.

that this mechanism of DNA recognition is distinct from the direct readout of hydrogen-bond donors and acceptors in the major groove, because different strings of nucleotides can adopt similar minor-groove shapes.

For this to be a means of recognizing information encoded in the DNA sequence, there must be particular nucleotide sequences that result in a narrow minor groove. Again, by searching the structural databases, Rohs *et al.*¹ found that short runs of adenine nucleotides (called A-tracts) have a strong tendency to induce a narrow minor groove. The special structural properties of A-tracts have been known for three decades, and have been principally associated with DNA-sequence-directed curvature³. Curvature of DNA requires a series of short A-tracts, spaced at the DNA helix repeat of 10 base pairs. What is new in Rohs and colleagues' analysis¹ is their demonstration that an isolated A-tract can influence DNA shape in such a way that it can serve as a site for specific recognition by DNA-binding proteins.

It's not only sequence-specific DNA-binding proteins that have figured out how to read the electrostatic potential of the minor groove. Rohs *et al.*¹ show that histone octamers, the protein complexes around which DNA wraps to form nucleosomes, also exploit this mechanism. By comparing nucleotide sequences of DNA segments known to form nucleosomes, they show that short A-tracts have a tendency to be periodically spaced throughout such sequences. They also find, from X-ray structures of nucleosomes, that arginines are often present in the minor groove where it narrows as a consequence of wrapping the DNA double helix around the histone octamer.

The ability to sense the variation in electrostatic potential in DNA may reveal how a protein could home in on its binding site in the genome without touching every nucleotide, as electrostatics is a through-space phenomenon.

A difficulty of applying the analysis presented in this study is that it depends on high-resolution three-dimensional structures of protein–DNA complexes — that's the necessary input to the Poisson–Boltzmann equation that was used by Rohs *et al.*¹ to calculate the electrostatic potential of the minor groove. If we could develop an experimental measure of the minor-groove potential and how it varies with sequence, it would be possible to read the human genome like a protein does, treating DNA as a molecule and not just a string of letters. ■

Tom Tullius is in the Department of Chemistry and the Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA. e-mail: tullius@bu.edu

1. Rohs, R. *et al.* *Nature* **461**, 1248–1253 (2009).
2. Seeman, N. C., Rosenberg, J. M. & Rich, A. *Proc. Natl Acad. Sci. USA* **73**, 804–808 (1976).
3. Wu, H.-M. & Crothers, D. M. *Nature* **308**, 509–513 (1984).

STATISTICAL PHYSICS

Swirled by light

Mark I. Dykman

A micrometre-sized particle immersed in a liquid can be trapped by light. An experiment shows that the trapping can be accompanied by a whirling whose direction can be reversed by changing the light intensity.

The past few decades have witnessed remarkable progress in the control and manipulation of tiny objects, including micrometre-sized particles, bacteria and DNA molecules. These advances are associated with the development of optical tweezers^{1,2} — instruments that use strongly focused laser light to trap objects and move them at will. However, the stiffness of optical tweezers is limited, and an optically trapped particle does not reside at the trap centre but rather wanders about it. Such wandering is a consequence of thermal fluctuations in the particle's surrounding liquid, and is a counterpart of the random motion experienced by free-floating particles — the Brownian motion explained by Albert Einstein in 1905.

Writing in *Physical Review E*, Sun *et al.*³ report that the trajectory of an optically trapped colloidal particle in water, although random, can display a complicated but identifiable pattern. The particle's motion is characterized by a circulatory bias — the streamlines that are obtained by averaging over an ensemble of trajectories exhibit circulation. The authors associate this circulation with a toroidal particle current and call it a Brownian vortex. Interestingly, they find that the overall streamline pattern and the direction of the circulation depend on the intensity of the laser light used to trap the particle, with the possible coexistence of counter-rotating rolls of toroidal current. (A spurious inward spiralling in the streamline pattern was addressed by

the authors in an independent study⁴.)

Sun and colleagues' observation³ of a macroscopic flux — in the form of streamline circulation — is both interesting in itself and a clear indication that the particle is not at thermal equilibrium. The flux was found in a stationary optical trap, in contrast to the directed motion of Brownian particles seen previously⁵ in non-stationary traps, in which a ratchet-like optical potential was periodically modulated in time.

The absence of macroscopic fluxes is a fundamental property of classical systems in thermal equilibrium. Such fluxes would dissipate energy through some form of friction, and would decay if there was no energy input from the outside. By contrast, non-equilibrium systems that gain energy from external sources should generally display fluxes. These fluxes can be stationary — a property compatible with the system as a whole being stationary (with no accumulation of particles, energy and so on) and with the probability distribution of its dynamical variables being independent of time. For optically trapped, transparent colloidal particles in thermal equilibrium, the stationary probability distribution has been studied in detail^{6,7}.

Non-equilibrium, flux-carrying systems can be divided into two classes. One comprises systems characterized by a stationary flux that is sustained by a periodic or constant external driving and does not require fluctuations (thermal or non-thermal) to persist.

A well-known physical example of such a system is a laser: once a laser starts radiating in response to external pumping, the primary effect of fluctuations is to reduce the coherence of the laser radiation — that is, the extent to which the emitted laser light waves are in step with one another decreases. By contrast, in systems belonging to the second class, external driving on its own cannot support a stationary flux; the very occurrence of the flux is due to fluctuations⁸. It is into this second category that the colloidal-particle system studied by Sun and colleagues³ falls.

An insight into the onset of a fluctuation-facilitated flux can be gained by looking at the fluctuation dynamics of a system. In the 1950s, Onsager and Machlup noticed⁹ that, for a colloidal particle in thermal equilibrium, the most likely trajectory from the equilibrium position to a given spatial point coincides with the time-reversed most likely path the particle would follow from that point. Therefore, most probably, the particle arrives at a point with a velocity opposite to the one with which it leaves the point. The implication is that there is no net flux. By contrast, the most probable trajectories of non-equilibrium systems, such as that studied by Sun *et al.*³, lack time-reversal symmetry, and this leads to the onset of a flux. That there is a difference between the most probable trajectories of a non-equilibrium system to and from a given point has been demonstrated both in analogue electrical circuits¹⁰ and experimentally¹¹.

Sun and colleagues' observation³ of a stationary flux and its intricate structure for an optically trapped colloidal particle provides insight into the physics of systems away from thermal equilibrium, demonstrating features in their dynamical behaviour that have no analogue in systems in thermal equilibrium. The trajectory-based approach is useful for gaining such insight because of its sensitivity to the dynamics of the system. In the context of optical tweezers and their applications, the experiment raises questions about the nature of the forces to which a particle is subjected in an optical trap and the reason the particle is not in thermal equilibrium. ■

Mark I. Dykman is in the Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA. e-mail: dykman@pa.msu.edu

1. Ashkin, A. *Phys. Rev. Lett.* **24**, 156–159 (1970).
2. Ashkin, A., Dziedzic, J., Bjorkholm, J. & Chu, S. *Opt. Lett.* **11**, 288–291 (1986).
3. Sun, B., Lin, J., Darby, E., Grosberg, A. Y. & Grier, D. G. *Phys. Rev. E* **80**, 010401 (2009).
4. <http://physics.nyu.edu/grierlab/sampledvortex>
5. Faucheux, L. P., Bourdieu, L. S., Kaplan, P. D. & Libchaber, A. J. *Phys. Rev. Lett.* **74**, 1504–1507 (1995).
6. Florin, E.-L., Pralle, A., Stelzer, E. & Hörber, J. *Appl. Phys. A* **66**, S75–S78 (1998).
7. McCann, L. I., Dykman, M. & Golding, B. *Nature* **402**, 785–787 (1999).
8. Tomita, K. & Tomita, H. *Prog. Theor. Phys.* **51**, 1731–1749 (1974).
9. Onsager, L. & Machlup, S. *Phys. Rev.* **91**, 1505–1512 (1953).
10. Luchinsky, D. G. & McClintock, P. V. E. *Nature* **389**, 463–466 (1997).
11. Chan, H. B., Dykman, M. I. & Stambaugh, C. *Phys. Rev. Lett.* **100**, 130602 (2008).

PROGRESS

Volatile accretion history of the terrestrial planets and dynamic implications

Francis Albarède¹

Accretion left the terrestrial planets depleted in volatile components. Here I examine evidence for the hypothesis that the Moon and the Earth were essentially dry immediately after the formation of the Moon—by a giant impact on the proto-Earth—and only much later gained volatiles through accretion of wet material delivered from beyond the asteroid belt. This view is supported by U–Pb and I–Xe chronologies, which show that water delivery peaked ~100 million years after the isolation of the Solar System. Introduction of water into the terrestrial mantle triggered plate tectonics, which may have been crucial for the emergence of life. This mechanism may also have worked for the young Venus, but seems to have failed for Mars.

Earth, Mars and Venus are three planets with very different histories. Active plate tectonics, the presence of a liquid ocean, and teeming life characterize the Earth. Mars has a tenuous atmosphere, shows no incontrovertible evidence of plate tectonics, lacks an abundant hydrosphere, and has so far evaded all attempts to reveal the presence of life. Venus is a dead inferno of carbon dioxide with clouds of sulphuric acid, and its solid surface seems to have been violently resurfaced some 600–900 million years (Myr) ago. At a time when so many satellites and telescopes are probing the Universe in search of habitable planets, it is worth questioning what made three planets of our Solar System so different from each other that only one of them harbours life.

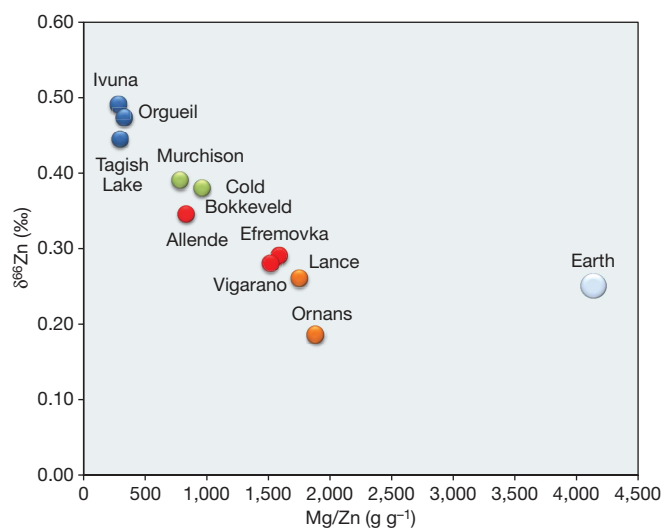


Figure 1 | Fractionation of Zn isotopes is incompatible with volatilization. The data⁸⁷ represent the $\delta^{66}\text{Zn}$ value, the relative deviation of the $^{66}\text{Zn}/^{64}\text{Zn}$ ratio in per mil with respect to a terrestrial standard, of each sample. Data for terrestrial samples are shown in pale grey; data for the different classes of carbonaceous chondrites are shown in dark blue (CI), green (CM), red (CV) and orange (CO). In the case of volatilization, preferential loss of the light ^{64}Zn isotope in the residue is expected to accompany the increase of the Mg/Zn ratio, as magnesium ($T_{50} = 1,336\text{ K}$) is much more refractory than zinc ($T_{50} = 726\text{ K}$). Because the opposite is observed, fractionation processes other than volatilization must be sought. T_{50} , at which 50% of the element is condensed¹³.

When life is the topic, the presence of liquid water immediately becomes an issue. The origin of water and its fate have also been central concerns of the planetary sciences, because water changes the rheological properties of planetary mantles and seems to be a prerequisite for plate tectonics to operate. A commonly held view is that the terrestrial ocean results from the outgassing of the Earth's mantle¹. Recent dynamic calculations, however, emphasize the hot temperatures next to the nascent Sun, and suggest that the planetary embryos from the inner Solar System remained essentially dry. It is only late in accretion history that perturbations of asteroid orbits in the inner Solar System by giant planets started to launch ice-rich material that delivered water to the planets in that part of the Solar System. A simplistic but efficient description of the debate on planetary water is that of 'dry' versus 'wet' accretion.

This Review will first show that the depletion of the volatile content of the Earth and other terrestrial planets is not due to loss by outgassing but to incomplete stepwise accretion of nebular material; this accretion was interrupted by energetic electromagnetic radiation that was emitted by the young Sun during its T Tauri phase and swept through the disk. The timing of volatile accretion, water in particular, with respect to the giant lunar impact, the formation of the terrestrial core, and its implications for the contrasting geodynamic regimes of the Earth, Mars, and Venus will be discussed.

Water in the deep Earth

The Earth contains water in different forms: a liquid ocean at the surface, various forms of ground water and, in the mantle, dense hydrous mineral phases and water dissolved in nominally anhydrous minerals such as olivine. Evidence for how much water the solid Earth contains is robust. In the absence of a gaseous phase, that is, at depths in excess of about 90 km, water behaves as an incompatible element with properties remarkably similar to those of cerium²: the $\text{H}_2\text{O}/\text{Ce}$ ratio is about 200 in the source of both mid-ocean ridge basalts and ocean island basalts, which are widely taken as melts from upper- and lower-mantle material, respectively. Given this strong constraint and reasonable assumptions on the degrees of melting prevailing during the formation of these melts, the amount of water held by the mantle is approximately equivalent to an oceanic mass, which translates into 150–350 p.p.m. H_2O in the mantle and a bulk terrestrial water concentration of 300–550 p.p.m. (refs 3, 4). The water content of the Archaean mantle is open to interpretation⁴,

¹Ecole Normale Supérieure de Lyon, Université Claude-Bernard Lyon 1, and CNRS, 46 allée d'Italie, 69007 Lyon, France.

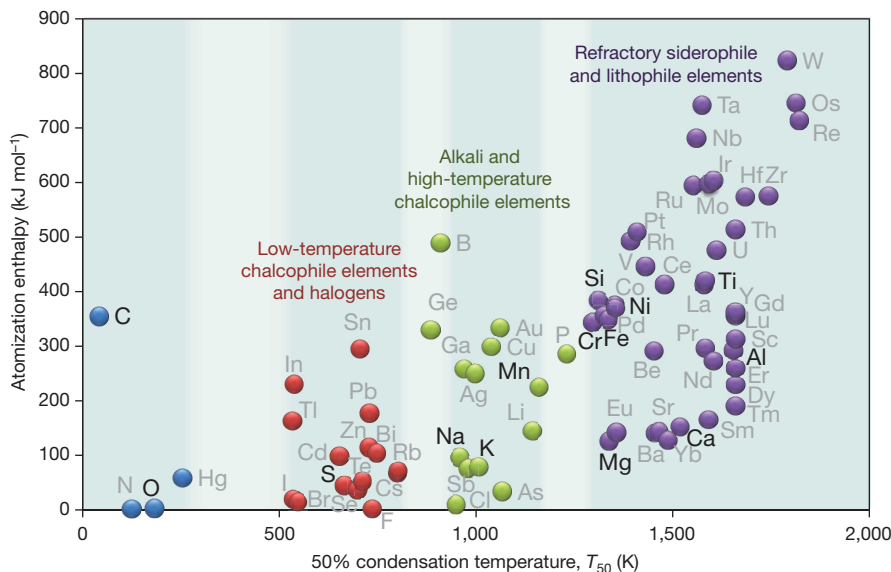


Figure 2 | Stepwise accretion of the elements on cooling of the solar nebula, shown as atomization enthalpy versus T_{50} . The two axes represent two different volatility scales, an intrinsic scale given by the atomization enthalpy²⁹, which is essentially equivalent to the mean bonding energy per atom in solids, and a scale dependent on the nebular chemistry, T_{50} (see Fig. 1 legend). Black letters, major elements; grey letters, minor and trace elements. The platinum group elements, plus Al, Ti, Zr and W, and most rare-earth elements and actinides, which condense first above 1,600 K, are preferentially found in refractory inclusions. The group of mildly refractory

lithophile elements, which comprise the major elements Si, Mg, Fe and Ca, accrete down to 1,300 K as metal, olivine and pyroxenes, and make up the bulk of the planetary mantle and core. These refractory elements are separated from the high- T chalcophile elements (As, Ga, Ge, Cu, Ag), chlorine, and the alkali elements (Li, Na, K, Rb and Cs), which condense between 1,150 and 850 K. From 750 to 530 K, the low- T chalcophile elements (Pb, Bi, Sn, Zn, Cd, S, Te) precipitate, followed by the truly volatile elements (N, C, H) and Hg. Each group appears to be separated from its neighbours by a temperature gap (highlighted in white).

but seems to have been lower than that of the modern mantle. As it is particularly difficult to ascertain whether the measured water contents of lavas reflect their contents at the time of eruption, the Ce contents of melt inclusions in olivine from the 2.7-Gyr-old komatiites from Belingwe, Zimbabwe⁵, and Alexo, Canada⁶, can be used as a proxy; they show that if the H_2O/Ce ratio has not significantly evolved since that time, the mantle at 2.7 Gyr ago contained ~ 150 p.p.m. H_2O for 50% melting and ~ 60 p.p.m. for 20% melting.

It has been suggested that much of the mantle's water may accumulate in the transition zone in wadsleyite^{7,8} and, therefore, that the water content of the mantle may not be homogeneous. This is true only if water is held back in the transition zone on subduction and not dragged down with its carrier minerals. Fractionation across the phase-change boundary requires that when wet material is transported in and out of the transition zone, the advective flux is compensated for by diffusion of hydrogen: at the temperatures prevailing in the sinking plates (600–1,200 °C), the diffusion coefficients of hydrogen in the major minerals are very small⁹ and the wet boundary layer appearing across the reaction zone very thin. Accumulation of water in wadsleyite is therefore of little significance to the overall budget of water in the mantle.

Volatile depletion in the terrestrial planets

Planets formed by accretion of solid material condensed from the solar nebula. The strong electromagnetic winds emitted by the early Sun swept away the nebular gas and condensation stopped well before the debris disk was cleared, a process that lasted less than a few Myr after the isolation of the Solar System¹⁰. Most of the material that eventually formed the Earth accreted well within the 'snow line', which is the virtual line beyond which water freezes out^{11,12}. A widely used scale of volatility is the temperature T_{50} at which 50% of the nebular inventory has been accreted to the solid phase. Let us consider two elements with widely different values of T_{50} , namely, the refractory element uranium ($T_{50} = 1,610$ K), and the more volatile element potassium ($T_{50} = 1,006$ K)¹³. The K/U ratio is a measure of the relative depletion of planetary volatile elements with respect to

refractory elements. Carbonaceous chondrites and the solar photosphere provide a coherent estimate of the initial K/U ratio of 60,000 (ref. 14). The terrestrial value of $\sim 10,000$ (refs 15, 16) indicates that the Earth lacks 85% of the nebular K inventory, while the deficit for the Moon ($K/U \approx 3,000$) is 95%. Martian meteorites that have been witnessed falling and therefore evaded long periods of terrestrial weathering show K/U ratios $< 20,000$ (ref. 17). Depletion of 92–98% is also inferred for Zn, Ag, As, Sb, Sn, Pb and, most importantly, S (refs 18, 19). Planets that lack such large fractions of moderately volatile elements cannot have been endowed with large amounts of water.

As shown by isotopic abundances of different elements, the volatile deficit in planetary material is not due to volatilization but to incomplete accretion²⁰. Loss of volatile elements due to impact or volatilization during accretion, whether it reflects planetary escape velocities or the molecular stampede of hydrogen escape known as hydrodynamic entrainment²¹, is expected to be mass-dependent, and the light isotopes should preferentially be enriched in the vapour phase. The homogenous stable isotope composition of K in planetary samples is inconsistent with preferential evaporation of the light isotopes from initially condensed (CI chondrite) material. Even more conspicuously, the Earth is depleted in the heavy isotopes of Zn by about 0.1‰ per atomic mass unit with respect to CI chondrites, as are the CV and CO chondrites, which are enriched in refractory elements (Fig. 1). This is exactly the opposite of what is expected from preferential loss of the light ^{64}Zn during volatilization. These observations contrast with evidence from lunar soils, which on volatilization by impacts become remarkably enriched in the heavy ^{66}Zn and ^{68}Zn isotopes by 3–4‰ per atomic mass unit²². The processes leading to reversed mass fractionation are currently not well understood, and may involve kinetic effects²³ or electromagnetic sorting (Hall effect) of the ionized fraction of the nebular gas²⁴. Incomplete accretion is also inferred from the terrestrial abundances of alkali elements, which have comparable volatilities and chemical properties over a broad range of atomic mass M . The depletions of K ($M = 39$, $T_{50} = 1,006$ K), Rb ($M = 85$, $T_{50} = 800$ K)²⁵ and Cs ($M = 133$,

$T_{50} = 800 \text{ K}$)²⁶ are of comparable magnitude, which also argues against loss by volatilization in a gravity field.

A robust consequence is that the Earth, like the other terrestrial planets, is depleted in volatile elements, not because these were lost by impact or vaporization, but because they did not accrete along with the more refractory elements in the first place.

The stepwise accretion of planetary material

In order to explain the volatile deficit, the idea being pursued here is that the nebular gas was blown off by the energetic radiation of the young Sun before temperature decreased enough for volatile elements to condense on the terrestrial planets. Thermodynamic calculations^{27,28} successfully predict the mineralogy and composition of planetary objects. Most elements condense within a narrow temperature interval. As a consequence, accretion takes place as a stepwise process and bulk accretion goes through plateaus when shifting from one group of elements to another with decreasing temperature.

The narrow temperature range of condensation of most elements is a common feature of thermodynamics calculations^{13,27,28} and reflects relatively simple properties of solid–gas equilibria. The temperature range ΔT over which elements condense is approximately $RT_{50}^2/\Delta H_{\text{at}}^i$, where R is the gas constant and ΔH_{at}^i is the atomization enthalpy²⁹. Energy of mineral formation from solid elements and energy of melting can be safely neglected with respect to the enormous energies involved in vaporization. For most elements, ΔT corresponds to a few tens of degrees, except for the alkali elements, Eu and Yb (150–200 K).

Intrinsic volatility is not the sole control of planetary abundances. Figure 2 is a plot of atomization enthalpy versus T_{50} . The highly refractory elements—that is, the platinum group elements, Al, Ti, Zr, W, and most rare-earth elements and actinides, which are all particularly abundant in chondrite refractory inclusions—condense above 1,600 K. They are followed by the dominant group of refractory lithophile elements (Si, P, and the alkali-earth and transition elements typical of chondrules), which condense down to 1,300 K. These refractory elements are separated from a group comprising the high- T chalcophile elements (As, Ga, Ge, Cu and Ag), chlorine, and the alkali elements (Li, Na, K, Rb and Cs), which accrete between 1,150 and 850 K. At lower temperatures (750 to 530 K), the group of low- T

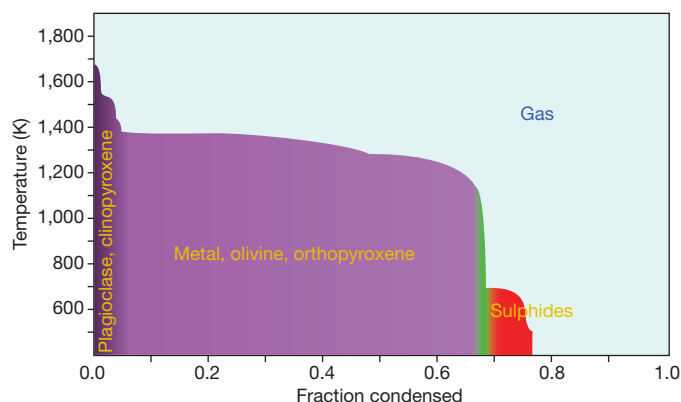


Figure 3 | Stepwise accretion of the elements upon cooling of the solar nebula, shown as fraction condensed versus temperature. Plotted is the fraction of the original condensable matter remaining in the solar nebula at a given temperature (redrawn with changes from ref. 30). The kinks in the accretion rate correspond to the temperature gaps in Fig. 2. After formation of the most abundant refractory phases (plagioclase and clinopyroxene), the bulk of the metallic core and the mantle silicates (olivine and orthopyroxene) is quickly removed from the nebular gas (purple). Alkali elements and high- T chalcophile elements precipitate next (green), but do not constitute a large fraction of the planetary material. Low- T chalcophile elements, sulphur and halogens come after (red), followed by volatile elements. Most elements accrete over a few tens of degrees, but over 150–250 K for the alkali elements, and Eu and Yb. A planet with a substantial deficit of K or Zn with respect to chondrites therefore cannot have accreted much, if any, water.

chalcophile elements (Pb, Bi, Sn, Zn, Cd, S and Te) and the rest of the halogens leave the vapour phase. A last group consists of the three most volatile elements (N, C and H) and Hg. Each group appears to be separated from its neighbouring groups by a temperature gap. The stepwise accretion of the planetary material is well illustrated in refs 30 and 31, which show that, at 1,200 K, 50–60% of the nebular inventory is still in the gas phase (Fig. 3). This reflects the combination of the elements into stoichiometric solid phases such as oxides, olivine, pyroxenes, feldspars and metal¹³, and, to a lesser extent, the intrinsic volatility of the elements themselves.

The energetic electromagnetic radiation of the young Sun swept away the nebular gas and stopped accretion while the nebular temperature in the neighbourhood of the terrestrial planets was still in the temperature range of alkali element condensation (800–1,000 K)^{30,32,33} (Fig. 4). This is the main cause of volatile element depletion in the inner Solar System. In contrast, in material accreted at the snow line and beyond, water and other volatile elements were plentiful. The ‘late veneer’ hypothesis³⁴, which holds that small amounts of chondritic material from the asteroidal belt or beyond were added to the Earth at an unspecified late stage, was devised to account for the present-day excess of highly siderophile elements, such as the platinum group elements, in the terrestrial mantle. The depletion of the most volatile elements discussed above places an upper limit on the proportion of low-temperature CI-like material of the order of 2–5%, which is still somewhat higher than the 0.3% proportion of carbonaceous chondrite-type (with ~10% water) material that is deemed⁴ necessary to account for terrestrial water.

Making planets and adding water

One model³⁵ of planetary growth and development may be summarized into three stages: (1) the settling of dust onto the equatorial disk of the planetary nebula and formation of kilometre-sized planetesimals, (2) the runaway growth of planetesimals into Mars-sized planetary embryos, and (3) collision of planetary embryos to form the planets with more or less their modern masses. How fast the nebular gas was

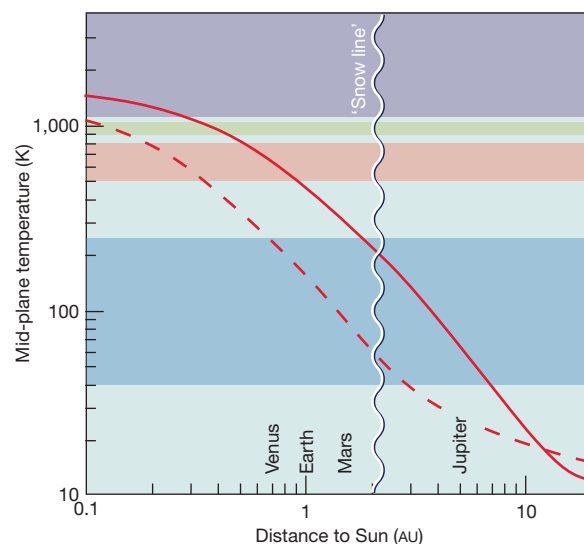


Figure 4 | The thermal structure of the planetary solar nebula: temperature at mid-plane of the nebular disk. Data are taken from ref. 88. Two models of temperature distribution are shown for two different values of the viscous dissipation parameter, α (viscosity \times sound velocity at mid-plane \times mean elevation above mid-plane) for a rate of solar accretion of 1% per Myr: $\alpha = 10^{-3}$ (solid line) and $\alpha = 10^{-1}$ (dashed line). The position of the wiggly ‘snow line’, which separates two domains, ice-free inwards and frosty outwards, nearly coincides with that of the asteroid belt¹². Colour coded areas correspond to the four groups of elements identified by their T_{50} values in Fig. 2. The depletion of volatiles in the inner Solar System is caused by the strong electromagnetic winds emitted by the young Sun, which swept away the nebular gas before accretion was complete.

blown off by energetic solar radiation is unclear: the delay is constrained by the short lifetime (<3 Myr) of the debris disk¹⁰ and by dynamical models, which require a reduction of planetary eccentricity by dynamic friction between planets, asteroids and planetary embryos³⁶. It is likely, however, that the planets may have accreted to a substantial fraction of their final mass before the disk was blown off.

Water delivery by comets³⁷ with a D/H ratio of 3×10^{-4} (refs 38, 39) seems incompatible with the terrestrial D/H ratio of the terrestrial ocean (1.5×10^{-4})⁴⁰. Carbonaceous chondrites represent an alternative source of terrestrial water^{1,41}. Orbits of planetary embryos across the Solar System during accretion have been computed⁴² and these authors concluded that the bulk of the water at present on Earth was carried by only a few of these objects, originally formed in the outer asteroid belt and accreted to the Earth during the waning stages of its formation. As long as the nebular gas was present and was absorbing electromagnetic radiation, the inner Solar System was too hot to allow for significant water condensation along with metal and silicates: only during the late stages of accretion had temperature declined enough for water delivery to be efficient. Dynamical models⁴³ suggest that early accretion of asteroid and planetary embryos was local, and therefore that the accreted material was largely dry up to the snow line and ice-rich beyond. It is also found that radial mixing and inward migration of wet high-eccentricity objects from the outer Solar System increased dramatically in the later phases of accretion. Water absorption in the hot inner Solar System has been proposed⁴⁴ but, as two-dimensional liquid films cannot form below the critical temperature of water, this process is unlikely to be significant.

The mechanism of water accretion depends on the relative timing of three important events—the segregation of the terrestrial core, the giant lunar impact, and the arrival of the late veneer (Fig. 5). The identical ^{182}W excesses found in the silicate fractions of the Earth and the Moon require that the giant impact took place at the earliest 60 Myr after the formation of calcium-aluminium-rich inclusions⁴⁵ unless the Hf/W ratios of the two planetary bodies also are similar, in which case this time can be reduced to ~ 30 Myr (refs 46, 47). A

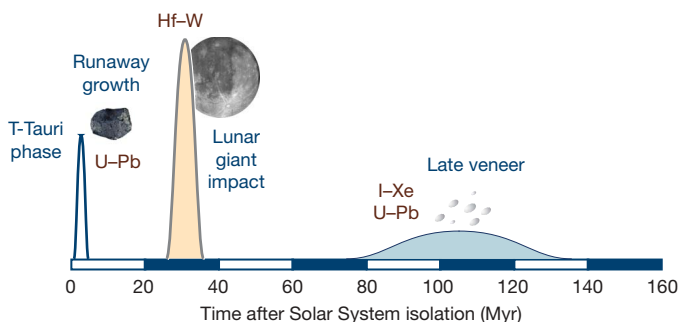


Figure 5 | A tentative chronology of the Earth's accretion. Chronometers shown in brown. Accretion of planetary material was interrupted by energetic electromagnetic radiation (T Tauri phase) sweeping across the disk within a few Myr of the isolation of the solar nebula. Runaway growth of planetesimals produces Mars-sized planetary embryos, which, collision after collision, form the planets with their modern masses. The last of these 'giant' collisions left material orbiting the Earth that later reassembled to form the Moon. The ^{182}Hf – ^{182}W chronometer dates metal–silicate separation. The identical abundance of radiogenic ^{182}W between the Earth and the Moon indicates that either the Moon formed after all the short-lived ^{182}Hf had disappeared (>60 Myr) or, rather, the Moon-forming impact and terrestrial core segregation took place simultaneously 30 Myr after isolation of the solar nebula. Addition of a late veneer of chondritic material coming from beyond 2.5 AU provides a strong explanation for the modern abundances of siderophile and volatile elements in the terrestrial mantle. This material also contained water and other volatile elements, which account for the origin of the terrestrial ocean. Such a model indicates that most of the terrestrial Pb and Xe was delivered by the asteroids that constituted the late veneer, and therefore that the young Pb–Pb and I–Xe ages of the Earth date, not the Earth, but events that affected the asteroids. It is suggested here that these events are those of the accretion to the Earth of the late veneer.

lunar impact with a wet Earth, that is, after substantial amounts of late veneer had been accreted, is difficult to reconcile with the remarkable depletion of volatiles in the lunar mantle^{30,48} with respect to the terrestrial mantle. The contrasting abundances of the stable (unradiogenic) Ne isotopes 20 and 22 (ref. 49) and of the stable Xe isotopes 124, 128 and 130 (ref. 50) between the terrestrial mantle, represented by basalts and CO_2 wells, and the terrestrial atmosphere also require that a substantial fraction of the volatiles was accreted after the last large-scale melting of the Earth. The invariant proportions of the stable Ar isotopes 36 and 38 in all terrestrial solids and gases^{51,52} indicates that isotope fractionation of Ne and Xe was not caused by hydrodynamic escape. Although the case has been made recently that the lunar mantle may locally hold traces of water⁵³, the depletion of the Moon in volatile elements, regardless of their atomic weights and therefore of gravitational loss, remains unexplained. However, the occasional closeness of planetary bodies with very different degrees of hydration and, because of lateral transfer, the highly variable water contents of impactors are strong features of dynamic simulations⁴³.

Because Pb is volatile, with a T_{50} value (~ 725 K; ref. 13) similar to that of Zn, the arrival of the late veneer can be dated by Pb–Pb chronology. Old feldspars and galenas indicate Pb–Pb ages of the Earth younger by 50–160 Myr relative to the age of the Solar System^{54,55} (Fig. 6). The delayed ingrowth of radiogenic Pb therefore is best explained by impacts of asteroids with very low, chondrite-like μ ($^{238}\text{U}/^{204}\text{Pb}$) ratios⁵⁶ onto a volatile-depleted, high- μ proto-Earth. The μ value of the lunar mantle attested to by Pb isotope systematics of Apollo samples is 200–600 times the ratio of the modern terrestrial mantle and 1000–10,000 times that of CI chondrites⁵⁷. Simple mass balance using the μ values suggests that well over 99% of terrestrial Pb was added by the late veneer. The young Pb–Pb age of the Earth therefore appears to be the mean age of the arrival of the late veneer material and does not represent a date in the history of primordial terrestrial Pb. Some U/Pb fractionation due to volatilization or to incomplete mergers ('hit-and-run' collisions) during the encounter of these asteroids with the Earth may account for the young Pb–Pb ages. Likewise, the apparent I–Xe and Pu–Xe 'ages of the Earth', which are bracketed between 50 and 110 Myr (refs 58–60), date the arrival on Earth of I and Xe, two highly volatile elements, and are

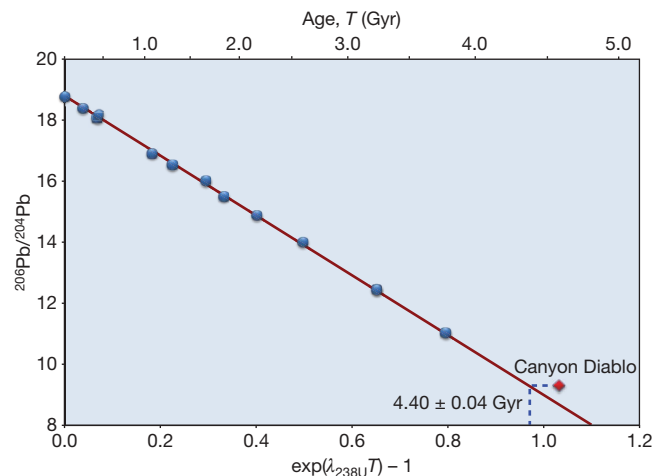


Figure 6 | Galena data and the young age of the Earth. Data are taken from ref. 54. The evolution of $^{206}\text{Pb}/^{204}\text{Pb}$ in conformable Pb deposits is linear in $\exp(\lambda_{238\text{U}} T) - 1$, where $\lambda_{238\text{U}}$ is the decay constant of ^{238}U , which indicates that the mantle source of galenas behaved as a closed system since the early geological ages. The primordial Pb of the Solar System, represented by the 4.56-Gyr-old Canyon Diablo troilite⁵⁹, is not on the galena trend, which suggests that terrestrial Pb is younger than that of this meteorite. This condition reflects that either Pb segregated into the core at a late stage⁵⁹ or, as suggested here, that terrestrial Pb was largely inherited from the late accretion of volatile-rich material (late veneer) to the Earth.

reasonably consistent with the Pb–Pb age. Hence, arrival of the late veneer about 100 ± 50 Myr after the collapse of the nebula is supported by independent sets of geochemical observations.

The nature of the wet, late-accreting icy bodies from beyond the snow line, either identified meteoritic or so-far unsampled material, is not understood^{1,41}. It is likely that any source of planetary material that significantly contributed to planetary accretion during the first 50 Myr of the Solar System is now exhausted. This makes the quest for a particular type of modern chondritic contributor intrinsically elusive.

Hydrating the mantle and setting off planetary dynamics

If water is a late addition to terrestrial accretion, the question arises of when and how modern mantle water was introduced into the solid Earth⁴. Solution in the primordial magma ocean (ingassing)⁶¹ requires a very high atmospheric temperature ($>800^\circ\text{C}$) to prevent mantle hydration. Foundering the hydrous crust of serpentine and other hydrous minerals⁶² of the magma ocean is an alternative mechanism for early deep Earth hydration. In both cases, however, crystallization of the magma ocean is expected to concentrate water, a very incompatible molecule, in the residual melt and therefore acts to remove most of the water initially present in the upper mantle.

Water and all the other incompatible elements are particularly depleted in the upper mantle for two main reasons. First, orogenic volcanism at converging plate boundaries removes most of the volatile and incompatible lithophile elements from the hydrous lithospheric plates. Later, mid-ocean-ridge magmas remove from the upper mantle most of what survived the subduction ‘filter’. In addition, it is a general feature that magma extraction under mid-ocean ridges and ocean islands is a rather shallow process (<100 km) that takes place during decompression of the mantle: the deeper into the mantle water and incompatible elements are buried, the less likely they are to become involved in melting and extraction, which is apparent in convection models dealing with geochemical tracers⁶³. Water is continuously lost to the deep mantle by subduction of a post-serpentine hydrous phase known as phase D^{64–66}. The presence of hydrous phases at the top of the lower mantle may actually account for stronger seismic attenuation at subduction zones than in the ambient mantle⁶⁷.

Water decreases the viscosity of major mantle minerals, notably olivine⁶⁸. It reduces the strain before lithospheric failure and therefore affects mantle dynamics⁶⁹. A strong reduction of mantle viscosity contingent on subtle olivine hydration is suspected to shift the elastic thickness of the lithosphere and to change mantle dynamics from a stagnant-lid regime (strong dry mantle) to a plate tectonic regime (soft wet mantle)⁷⁰.

The varying fates of water on the different terrestrial planets provide a relatively simple explanation of their contrasting evolution. Admitting that water and Ce have similar residence times of 6–8 Gyr (ref. 71) in the terrestrial mantle, $\sim 50\%$ of the ocean must have been subducted into the mantle after 4.5 Gyr, which is precisely what the widely held view of ‘one ocean inside for one ocean outside’ is predicting.

The status of water on Venus can be explained by conditions opposite to those on the Earth: hot surface vapour quickly reacts with the crust and is rapidly transported downwards because the more hydrous the mantle, the faster its convection. The loss of water from the surface is largely controlled by its reactivity and therefore by surface temperature. On both planets, the depletion of the upper mantle acts to hold water back in the lower mantle. The unique hypsometric maximum on Venus contrasts, however, with the bimodal distribution of elevations on Earth⁷² and demonstrates its lack of true continents. Evidence of plate tectonics, if it existed in the past, was erased by resurfacing ~ 600 – 900 Myr ago^{73,74}. Although some water is currently lost from the atmosphere of Venus⁷⁵, the relative abundances of atmospheric rare gases argues against hydrodynamic escape in the past⁷⁶. If Venus and the Earth initially had oceans of similar sizes, an alternative may be found to atmospheric loss on Venus: water may simply have been dragged into the mantle by early

plate tectonics. Eventually, this may become the fate of the terrestrial oceans and continents as well, once enough water has been introduced into the mantle, making it softer than it is today.

It has been argued⁷⁷ that Martian water was delivered by asteroids and comets from beyond 2.5 AU. Also, it was suggested that Mars’ interior is dry because, although the planet acquired a late volatile-rich veneer, it did not get folded into the mantle⁷⁸. A dry mantle explains the early demise of plate tectonics on Mars. In contrast to the Earth and Venus, Mars lost its atmosphere to space by hydrodynamic escape during the first few hundred Myr after accretion⁷⁹. Fractionation of the atmospheric $^{38}\text{Ar}/^{36}\text{Ar}$ ratio attests to hydrodynamic escape⁷⁶. The weak gravity field due to the small planet radius has a triple effect: (1) escape from the atmosphere is much easier on Mars than on Earth; (2) because the pressure gradient is about three times weaker on Mars, the phase changes giving rise to the viscosity jump of the terrestrial transition zone are shifted to the bottom of the Martian mantle, thereby preventing significant accumulation of water at depth; and (3) the bending of plates and, hence, the onset of plate tectonics is made more difficult. It has been suggested that the mantle source of Martian shergottites could be wet^{80,81}, but this is only possible if these rocks are younger than a few hundred Myr, a point that was recently challenged on the basis of Pb–Pb ages in excess of 4 Gyr (ref. 82).

Contrasting dynamic regimes, and, therefore, the amount of water present in planetary mantles, have diverse effects on the dynamo: a plate tectonic regime allows heat to be evacuated more efficiently from the core than a stagnant-lid regime⁸³. The Earth, and possibly early Venus, illustrates the former case with active plate tectonics. In the latter case, the dynamo chokes, the magnetosphere vanishes, and the atmosphere gets eroded by solar wind: this was probably the fate of Mars, initially triggered by a water-poor mantle and the weak gravity field of the planet, and may become the fate of Venus.

Future directions

Beyond the fascinating question of how the Earth formed and acquired its hydrosphere lies the question of the origin of life. Associating life on Earth with plate tectonics and the presence of continents, all of them being contingent on the presence of water, is tantalizing and justifies the ‘follow the water’ motto of space agencies. Although the reducing environments created on reduction of water by the magma ocean may have jump-started biological activity⁸⁴, it is not clear that a ‘water world’ can provide a steady source of nutrients and sustain life⁸⁵. Both the timing of water delivery (relevant to the onset of plate tectonics) and the amount of water (for the persistence of a shallow ocean with subaerial reliefs) need to be understood. The detailed chemical reactions at work during the giant lunar impact⁸⁶ and even its timing remain poorly understood. New dynamical models of accretion will need to deal with radial transfer across the nebula and with hit-and-run impacts that severely affect planetary chemistry. The past decade witnessed major conceptual changes in our understanding of the early history of the terrestrial planets, and more surprises seem to be ahead of us.

1. Drake, M. J. & Righter, K. Determining the composition of the Earth. *Nature* **416**, 39–44 (2002).
2. Michael, P. Regionally distinctive sources of depleted MORB: evidence from trace elements and H_2O . *Earth Planet. Sci. Lett.* **131**, 301–320 (1995).
3. Brackets the geochemical behaviour of water during magma generation between that of rare-earth elements lanthanum and neodymium.
4. Saal, A. E., Hauri, E. H., Langmuir, C. H. & Perfit, M. R. Vapour undersaturation in primitive mid-ocean-ridge basalt and the volatile content of Earth’s upper mantle. *Nature* **419**, 451–455 (2002).
5. Marty, B. & Yokochi, R. in *Water in Nominally Anhydrous Minerals* (eds Keppler, H. & Smyth, J. R.) 421–450 (Rev. Mineral. Geochem. 62, Mineral. Soc. Am., 2006).
6. McDonough, W. F. & Ireland, T. R. Intraplate origin of komatiites inferred from trace-elements in glass inclusions. *Nature* **365**, 432–434 (1993).
7. Lahaye, Y., Barnes, S. J., Frick, L. R. & Lambert, D. D. Re-Os isotopic study of komatiitic volcanism and magmatic sulfide formation in the southern Abitibi greenstone belt, Ontario, Canada. *Can. Mineral.* **39**, 473–490 (2001).

7. Bercovici, D. & Karato, S. I. Whole-mantle convection and the transition-zone water filter. *Nature* **425**, 39–44 (2003).
8. Bolfan-Casanova, N., Keppler, H. & Rubie, D. C. Water partitioning between nominally anhydrous minerals in the MgO–SiO₂–H₂O system up to 24 GPa: implications for the distribution of water in the Earth's mantle. *Earth Planet. Sci. Lett.* **182**, 209–221 (2000).
9. Ingrin, J. & Blanchard, M. in *Water in Nominally Anhydrous Minerals* (eds Keppler, H. & Smyth, J. R.) 291–320 (Rev. Mineral. Geochem. 62, Mineral. Soc. Am., 2006).
10. Wyatt, M. C. Evolution of debris disks. *Annu. Rev. Astron. Astrophys.* **46**, 339–383 (2008).
11. Stevenson, D. J. & Lunine, J. I. Rapid formation of Jupiter by diffusive redistribution of water vapor in the solar nebula. *Icarus* **75**, 146–155 (1988).
12. Ciesla, F. J. & Cuzzi, J. N. The evolution of the water distribution in a viscous protoplanetary disk. *Icarus* **181**, 178–204 (2006).
13. Lodders, K. Solar system abundances and condensation temperatures of the elements. *Astrophys. J.* **591**, 1220–1247 (2003).
14. Taylor, S. R. in *Origin of the Moon* (eds Hartmann, W. K., Phillips, R. J. & Taylor, G. J.) 125–143 (Lunar Planet. Inst., 1984).
15. Jochum, K. P., Hofmann, A. W., Ito, E., Seufert, H. M. & White, W. M. K. U and Th in mid-ocean ridge basalt glasses and heat production, K/U and K/Rb in the mantle. *Nature* **306**, 431–436 (1983).
16. Wasserburg, G. J., MacDonald, G. J. F., Hoyle, F. & Fowler, W. A. Relative contributions of uranium, thorium, and potassium to heat production in the Earth. *Science* **143**, 465–467 (1964).
17. Lodders, K. A survey of shergottite, nakhlite and chassigny meteorites whole-rock compositions. *Meteorit. Planet. Sci.* **33**, A183–A190 (1998).
18. Dreibus, G. & Palme, H. Cosmochemical constraints on the sulfur content in the Earth's core. *Geochim. Cosmochim. Acta* **60**, 1125–1130 (1996).
19. O'Neill, H. & Palme, H. in *The Earth's Mantle, Composition, Structure and Evolution* (ed. Jackson, I.) 3–126 (Cambridge Univ. Press, 1998).
20. Humayun, M. & Clayton, R. N. Potassium isotope cosmochemistry: genetic implications of volatile element depletion. *Geochim. Cosmochim. Acta* **59**, 2131–2148 (1995).
- This shows that the lack of fractionation between the isotopes of potassium among planetary bodies argues against major devolatilization of planetary bodies by impacts.**
21. Hunten, D. M., Pepin, R. O. & Walker, J. C. G. Mass fractionation in hydrodynamic escape. *Icarus* **69**, 532–549 (1987).
22. Moynier, F., Albarède, F. & Herzog, G. Isotopic fractionation of Zn, Cu and Fe in lunar materials. *Geochim. Cosmochim. Acta* **70**, 6103–6117 (2006).
23. Richter, F. M. Timescales determining the degree of kinetic isotope fractionation by evaporation and condensation. *Geochim. Cosmochim. Acta* **68**, 4971–4992 (2004).
24. Moynier, F. *et al.* Europium isotopic variations in Allende CAIs and the nature of mass-dependent fractionation in the solar nebula. *Geochim. Cosmochim. Acta* **70**, 4287–4294 (2006).
25. Gast, P. W. Limitations on the composition of the upper mantle. *J. Geophys. Res.* **65**, 1287–1297 (1960).
26. McDonough, W. F., Sun, S. S., Ringwood, A. E., Jagoutz, E. & Hofmann, A. W. Potassium, rubidium, and cesium in the Earth and Moon and the evolution of the mantle of the Earth. *Geochim. Cosmochim. Acta* **56**, 1001–1012 (1992).
27. Grossman, L. Condensation in the primitive solar nebula. *Geochim. Cosmochim. Acta* **36**, 597–619 (1972).
28. Larimer, J. W. Chemical fractionations in meteorites – I. Condensation of the elements. *Geochim. Cosmochim. Acta* **31**, 1215–1238 (1967).
29. Wulfsberg, G. *Inorganic Chemistry* (University Science Books, 2000).
30. Ganapathy, R. & Anders, E. Bulk compositions of the Moon and Earth, estimated from meteorites. *Proc. Lunar Planet. Sci. Conf.* **5**, 1181–1206 (1974).
31. Davis, A. M. & Richter, F. M. in *Treatise on Geochemistry* Vol. 2 (ed. Davis, A. M.) 407–430 (Elsevier, 2006).
32. Wänke, H. Constitution of terrestrial planets. *Phil. Trans. R. Soc. Lond. B* **303**, 287–302 (1981).
33. Ebel, D. S. & Grossman, L. Condensation in dust-enriched systems. *Geochim. Cosmochim. Acta* **64**, 339–366 (2000).
34. Chou, C. L. Fractionation of siderophile elements in the Earth's upper mantle. *Proc. Lunar Planet. Sci. Conf.* **9**, 219–230 (1978).
35. Wetherill, G. W. in *Origin of the Moon* (eds Hartmann, W. K., Phillips, R. J. & Taylor, G. J.) 519–551 (Lunar Planet. Inst., 1986).
36. O'Brien, D. P., Morbidelli, A. & Levison, H. F. Terrestrial planet formation with strong dynamical friction. *Icarus* **184**, 39–58 (2006).
- A remarkable attempt at understanding the history of water transfer across the Solar System during planetary accretion.**
37. Owen, T. & Bar-Nun, A. Comets, impacts, and atmospheres. *Icarus* **116**, 215–226 (1995).
38. Bockelée-Morvan, D. *et al.* Deuterated water in comet C/1996 B2 (Hyakutake) and its implications for the origin of comets. *Icarus* **133**, 147–162 (1998).
39. Huebner, W. Composition of comets: observations and models. *Earth Moon Planets* **89**, 179–195 (2000).
40. Lécuyer, C., Gillet, P. & Robert, F. The hydrogen isotope composition of seawater and the global water cycle. *Chem. Geol.* **145**, 249–261 (1998).
41. Robert, F. The origin of water on Earth. *Science* **293**, 1056–1058 (2001).
42. Morbidelli, A. *et al.* Source regions and time scales for the delivery of water to the Earth. *Meteorit. Planet. Sci.* **35**, 1309–1320 (2000).
- A breakthrough paper on accretion dynamics, demonstrating the prominence of water delivery to the inner Solar System from Jupiter and beyond.**
43. Raymond, S. N., Quinn, T. & Lunine, J. I. High-resolution simulations of the final assembly of Earth-like planets I. Terrestrial accretion and dynamics. *Icarus* **183**, 265–282 (2006).
44. Muralidharan, K., Deymier, P., Stimpfl, M., de Leeuw, N. H. & Drake, M. J. Origin of water in the inner Solar System: a kinetic Monte Carlo study of water adsorption on forsterite. *Icarus* **198**, 400–407 (2008).
45. Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals. *Nature* **450**, 1206–1209 (2007).
- An outstanding ¹⁸²Hf–¹⁸²W data set critical for the discussion of the early lunar chronology.**
46. Kleine, T., Münker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf–W chronometry. *Nature* **418**, 952–955 (2002).
47. Yin, Q. *et al.* A short timescale for terrestrial planet formation from Hf–W chronometry of meteorites. *Nature* **418**, 949–952 (2002).
48. Warren, P. H. in *Treatise on Geochemistry* Vol. 1 (ed. Davis, A. M.) 559–599 (Elsevier, 2005).
49. Honda, M., McDougall, I., Patterson, D. B., Dougeris, A. & Clague, D. Possible solar noble-gas component in Hawaiian basalts. *Nature* **349**, 149–151 (1991).
50. Caffee, M. W. *et al.* Primordial noble gases from Earth's mantle: identification of a primitive volatile component. *Science* **285**, 2115–2118 (1999).
- The first incontrovertible evidence that the Earth's atmosphere and mantle have different abundances of stable xenon isotopes.**
51. Kunz, J. Is there solar argon in the Earth's mantle? *Nature* **399**, 649–650 (1999).
52. Trieloff, M., Kunz, J. & Allègre, C. J. Noble gas systematics of the Reunion mantle plume source and the origin of primordial noble gases in Earth's mantle. *Earth Planet. Sci. Lett.* **200**, 297–313 (2002).
53. Saal, A. E. *et al.* Volatile content of lunar volcanic glasses and the presence of water in the Moon's interior. *Nature* **454**, 192–196 (2008).
54. Albarede, F. & Juteau, M. Unscrambling the lead model ages. *Geochim. Cosmochim. Acta* **48**, 207–212 (1984).
55. Galer, S. J. G. & Goldstein, S. L. in *Earth Processes: Reading the Isotopic Code* (eds Basu, A. & Hart, S. R.) 75–98 (Geophys. Monogr. 95, AGU, 1996).
56. Palme, H. & O'Neill, H. S. C. in *Treatise on Geochemistry* Vol. 2 (ed. Carlson, R. W.) 1–38 (Elsevier, 2005).
57. Premo, W. R., Tatsumoto, M., Misawa, K., Nakamura, N. & Kita, N. T. in *Planetary Petrology and Geochemistry: The Lawrence A. Taylor 60th Birthday volume* (eds Snyder, G. A., Neal, C. R. & Ernst, W. G.) 207–240 (Bellwether, 1999).
58. Pepin, R. O. & Phinney, D. The formation interval of the Earth. *Lunar Planet. Sci.* **VII**, 683–684 (1976).
59. Allegre, C. J., Manhès, G. & Gopel, C. The age of the Earth. *Geochim. Cosmochim. Acta* **59**, 1445–1456 (1995).
60. Ozima, M. & Podosek, F. A. Formation age of Earth from ¹²⁹I/¹²⁷I and ²⁴⁴Pu/²³⁸U systematics and the missing Xe. *J. Geophys. Res.* **B 104**, 25493–25499 (1999).
- A reference paper on how short-lived radioactivities based on Xe isotopes constrain the age of the Earth and of its atmosphere.**
61. Harper, C. L. & Jacobsen, S. B. Noble gases and Earth's accretion. *Science* **273**, 1814–1818 (1996).
62. Boyet, M. *et al.* ¹⁴²Nd evidence for early Earth differentiation. *Earth Planet. Sci. Lett.* **214**, 427–442 (2003).
63. Gurnis, M. & Davies, G. F. The effect of depth-dependent viscosity on convective mixing in the mantle and the possible survival of primitive mantle. *Geophys. Res. Lett.* **13**, 541–544 (1986).
64. Shieh, S. R., Mao, H.-k., Hemley, R. J. & Ming, L. C. Decomposition of phase D in the lower mantle and the fate of dense hydrous silicates in subducting slabs. *Earth Planet. Sci. Lett.* **159**, 13–23 (1998).
65. Kawamoto, T. in *Water in Nominally Anhydrous Minerals* (eds Keppler, H. & Smyth, J. R.) 273–289 (Rev. Mineral. Geochem. 62, Mineral. Soc. Am., 2006).
66. Smyth, J. R. in *Water in Nominally Anhydrous Minerals* 85–115 (Rev. Mineral. Geochem. 62, Mineral. Soc. Am., 2006).
67. Lawrence, J. F. & Wyssession, M. E. in *Earth's Deep Water Cycle* (eds Jacobsen, S. D. & van der Lee, S.) 251–260 (AGU Monograph 168, Am. Geophys. Un., 2006).
68. Hirth, G. & Kohlstedt, D. L. Water in the oceanic upper mantle: implications for rheology, melt extraction and the evolution of the lithosphere. *Earth Planet. Sci. Lett.* **144**, 93–108 (1996).
69. Kohlstedt, D. L., Evans, B. & Mackwell, S. J. Strength of the lithosphere – constraints imposed by laboratory experiments. *J. Geophys. Res.* **B 100**, 17587–17602 (1995).
70. O'Neill, C., Jellinek, A. M. & Lenardic, A. Conditions for the onset of plate tectonics on terrestrial planets and moons. *Earth Planet. Sci. Lett.* **261**, 20–32 (2007).
71. Albarède, F. in *Earth's Deep Mantle: Structure, Composition, and Evolution* (eds van der Hilst, R. D., Bass, J., Matas, J. & Trampert, J.) 27–46 (Geophys. Monogr. 160, Am. Geophys. Union, 2005).
72. Rosenblatt, P., Pinet, P. C. & Thouvenot, E. Comparative hypsometric analysis of Earth and Venus. *Geophys. Res. Lett.* **21**, 465–468 (1994).
73. Hauck, S. A., Phillips, R. J. & Price, M. H. Venus: crater distribution and plains resurfacing models. *J. Geophys. Res.* **E 103**, 13635–13642 (1998).
74. Phillips, R. J. *et al.* Impact craters and Venus resurfacing history. *J. Geophys. Res.* **97**, 15923–15948 (1992).
75. Barabash, S. *et al.* The loss of ions from Venus through the plasma wake. *Nature* **450**, 650–653 (2007).

76. Pepin, R. O. On the origin and early evolution of terrestrial planet atmospheres and meteoritic volatiles. *Icarus* **92**, 2–79 (1991).
 77. Lunine, J. I., Chambers, J., Morbidelli, A. & Leshin, L. A. The origin of water on Mars. *Icarus* **165**, 1–8 (2003).
 78. Carr, M. H. & Wänke, H. Earth and Mars: water inventories as clues to accretional histories. *Icarus* **98**, 61–71 (1992).
 79. Jakosky, B. M. & Phillips, R. J. Mars' volatile and climate history. *Nature* **412**, 237–244 (2001).
 80. McSween, H. Y. Jr *et al.* Geochemical evidence for magmatic water within Mars from pyroxenes in the Shergotty meteorite. *Nature* **409**, 487–490 (2001).
 81. Medard, E. & Grove, T. L. Early hydrous melting and degassing of the Martian interior. *J. Geophys. Res. Planets* **111**, doi:10.1029/2006JE002742 (2006).
 82. Bouvier, A., Blichert-Toft, J., Vervoort, J. D. & Albarede, F. The age of SNC meteorites and the antiquity of the Martian surface. *Earth Planet. Sci. Lett.* **240**, 221–233 (2005).
 83. Nimmo, F. & Stevenson, D. J. Influence of early plate tectonics on the thermal evolution and magnetic field of Mars. *J. Geophys. Res.* **105**, 11969–11979 (2000).
 84. Sleep, N. H., Meibom, A., Fridriksson, T., Coleman, R. G. & Bird, D. K. H₂-rich fluids from serpentinization: geochemical and biotic implications. *Proc. Natl Acad. Sci. USA* **101**, 12818–12823 (2004).
 85. Albarede, F. & Blichert-Toft, J. in *Origins of Life: Self-Organization and/or Biological Evolution?* (eds Gerin, M. & Maurel, M. C.) 1–12 (EDP Sciences, Paris, 2009).
 86. Pahlevan, K. & Stevenson, D. J. Equilibration in the aftermath of the lunar-forming giant impact. *Earth Planet. Sci. Lett.* **262**, 438–449 (2007).
 87. Luck, J.-M., Othman, D. B. & Albarede, F. Zn and Cu isotopic variations in chondrites and iron meteorites: early solar nebula reservoirs and parent-body processes. *Geochim. Cosmochim. Acta* **69**, 5351–5363 (2005).
- The strange behaviour of two non-traditional stable isotope systems and their bearing on planetary accretion.**
88. Fromang, S., Terquem, C. & Balbus, S. The ionization fraction in alpha models of protoplanetary disks. *Mon. Not. R. Astron. Soc.* **329**, 18–28 (2002).
 89. Tatsumoto, M., Knight, R. J. & Allegre, C. J. Time differences in formation of meteorites as determined from ratio of lead-207 to lead-206. *Science* **180**, 1279–1283 (1973).
- Acknowledgements** I am grateful to J. Blichert-Toft, S. Labrosse and H. Ohmoto for suggestions on the manuscript. Reviews by A. Morbidelli, M. Humayun and M. Drake were particularly helpful. Thanks to A. Levander and C.-T. Lee, I was able to spend enough quiet time at Rice University to bring this work to completion. This work was supported by the Agence Nationale de la Recherche and the Programme National de Planétologie (INSU-CEA).
- Author Information** Reprints and permissions information is available at www.nature.com/reprints. Correspondence should be addressed to the author (albarede@ens-lyon.fr).

REVIEWS

What recent ribosome structures have revealed about the mechanism of translation

T. Martin Schmeing¹ & V. Ramakrishnan¹

The high-resolution structures of ribosomal subunits published in 2000 have revolutionized the field of protein translation. They facilitated the determination and interpretation of functional complexes of the ribosome by crystallography and electron microscopy. Knowledge of the precise positions of residues in the ribosome in various states has facilitated increasingly sophisticated biochemical and genetic experiments, as well as the use of new methods such as single-molecule kinetics. In this review, we discuss how the interaction between structural and functional studies over the last decade has led to a deeper understanding of the complex mechanisms underlying translation.

The ribosome is the large ribonucleoprotein particle that synthesizes proteins in all cells, using messenger RNA as the template and aminoacyl-transfer RNAs as substrates. Ribosomes from bacteria consist of a large (50S) and a small (30S) subunit, which together compose the 2.5-megadalton 70S ribosome; their eukaryotic counterparts are the 60S and 40S subunits and the 80S ribosome. The 50S subunit consists of 23S RNA (~2,900 nucleotides), 5S RNA (~120 nucleotides) and about 30 proteins; the 30S subunit consists of 16S RNA (~1,500 nucleotides) and about 20 proteins. In addition, several protein factors act on the ribosome at various stages of translation. In this review, we focus mainly on structural and mechanistic insights into bacterial translation obtained in the last few years. A previous review deals more extensively with earlier work¹.

The essentially complete atomic structures of an archaeal 50S subunit from *Haloarcula marismortui*² and a bacterial 30S subunit from *Thermus thermophilus*³ published in 2000 were the basis for the phasing and/or molecular interpretation of every subsequent structure of the ribosome or its subunits. Such structures include low-resolution structures of the 70S ribosome by crystallography⁴ or cryoelectron microscopy (cryoEM)⁵, the structure of a bacterial 50S subunit⁶, and more recent high-resolution structures of the 70S ribosome^{7,8}. Finally, mobile elements of the 50S subunit such as the L1 or L7/L12 stalks that are partly or completely disordered in most high-resolution structures of the ribosome or the 50S subunit have been solved in isolation^{9,10}.

The basic architecture of the ribosome is shown in Fig. 1. The interface between the two subunits consists mainly of RNA. The mRNA binds in a cleft between the 'head' and 'body' of the 30S subunit, where its codons interact with the anticodons of tRNA. There are three binding sites for tRNA: the A site that binds the incoming aminoacyl-tRNA, the P site that holds the peptidyl-tRNA attached to the nascent polypeptide chain, and the E (exit) site to which the deacylated P-site tRNA moves after peptide-bond formation before its ejection from the ribosome. In the 50S subunit, the 3' ends of P- and A-site tRNAs are in close proximity in the peptidyl-transferase centre (PTC), whereas the 3' end of the E-site tRNA is ~50 Å away from the PTC.

Initiation

Bacterial translation can be roughly divided into three main stages, initiation, elongation and termination (Fig. 2; a movie of the process

can be seen at http://www.mrc-lmb.cam.ac.uk/ribo/homepage/movies/translation_bacterial.mov). Initiation requires the ribosome to position the initiator fMet-tRNA^{fMet} over the start codon of mRNA in the P site. In bacteria, the ribosome is positioned in the vicinity of the start codon by base pairing between the 3' end of 16S RNA and an approximately complementary sequence just upstream of the mRNA start codon, called the Shine-Dalgarno sequence. The precise positioning of the start codon in the P site requires the binding of a special initiator fMet-tRNA^{fMet} and three initiation factors, IF1–3. However, exactly how the correct tRNA is selected remains unclear, as are the roles of the various factors.

A probable first step in initiation is the binding of IF3 to the 30S that has been split from the 50S by ribosome recycling factor RRF and elongation factor G (EF-G) after translational termination (see Fig. 2 and the termination section later). This binding stimulates release of the mRNA and deacylated tRNA, leftover from the previous round of translation, from the 30S and prevents the large subunit from re-associating^{11,12}. The binding of the 30S–IF3 complex to mRNA, IF1, IF2 and initiator tRNA results in the 30S initiation complex (30S-IC). IF2, a GTPase, promotes subunit joining to form the 70S initiation complex (70S-IC), which is accompanied by IF3 release^{13–15}. After GTP hydrolysis and phosphate release from IF2 (refs 16, 17), fMet-tRNA^{fMet} moves into the PTC, readying the ribosome for elongation.

The mechanism of initiation is still unclear, owing to a paucity of structural data. There has been little progress towards high-resolution structures of initiation complexes since the structure of IF1 bound to a 30S subunit¹⁸. However, recent cryoEM studies have visualized both 30S and 70S initiation complexes. In a 30S-IC (ref. 19), which unfortunately did not contain IF3, IF2 stretches across the subunit interface of the 30S, contacting the acceptor end of fMet-tRNA^{fMet} with its carboxy terminus. The anticodon stem and elbow are shifted towards the E site, resulting in a '30S P/I state'. IF1 is visible in the A site, but does not contact IF2. After subunit joining, the G domain of IF2 interacts with the GTPase centre of the large subunit²⁰. It maintains its contacts with fMet-tRNA^{fMet}, which has shifted up out of plane from the 30S P/I state to a 70S P/I state, and seems to make a direct contact with IF1 in the 70S-IC. The 30S subunit is rotated relative to the 50S by ~4° anticlockwise, similar to the ratcheting seen during translocation²¹.

In the structure of 70S-mRNA–fMet-tRNA^{fMet}–IF2–GDPCP²², IF2 is still bound to the GTPase centre, but has lost contact with fMet-tRNA^{fMet}, now in the PTC in the canonical P/P state. The

¹MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK.

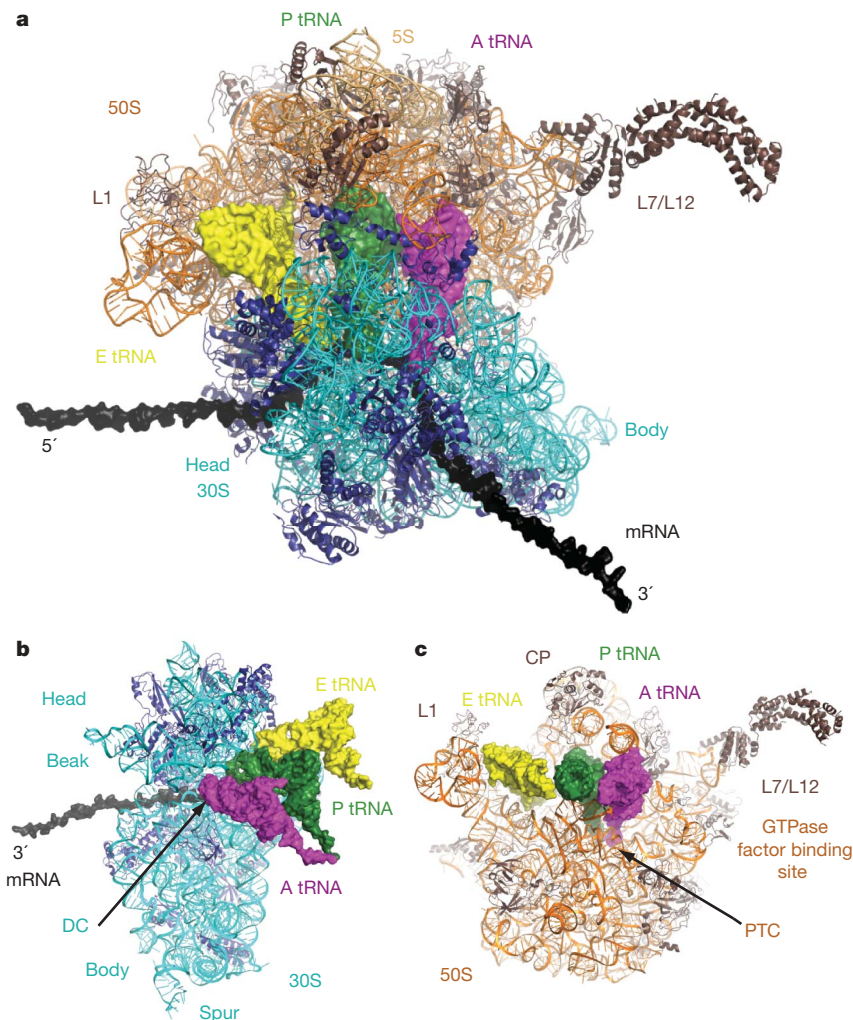


Figure 1 | Structure of the ribosome. **a**, 'Top' view of the 70S ribosome with mRNA and A-, P- and E-site tRNAs. **b**, **c**, Exploded view of the 30S subunit (**b**) and 50S subunit (**c**). The structure of the L7/L12 arm¹⁰ was fit onto the

70S ribosome⁶⁹, with mRNA elongated by modelling. This and all other figures were made with Pymol (Delano Scientific) and Photoshop (Adobe).

authors have suggested that this conformation represents the state after GTP hydrolysis before P_i release. Alternatively, another group has suggested that it is the result of the absence of IF1 and IF3 (ref. 23). The 70S complex with the GDP state of IF2 has the 30S subunit returned to the un-ratcheted state and IF2 largely separated from the GTPase centre, ready to dissociate from a properly initiated 70S ribosome²². Single-molecule fluorescence resonance energy transfer (FRET) studies show that this subunit rotation, which readies the ribosome for elongation, requires GTP hydrolysis²⁴, thus supporting a direct role for the GTPase activity of IF2 in initiation, which has been in dispute^{16,25}.

The elongation cycle

The elongation cycle consists of the steps involved in sequentially adding amino acids to the polypeptide chain (Fig. 2). At the beginning of the cycle, the ribosome contains a peptidyl-tRNA with a nascent polypeptide chain in the P site and an empty A site. During decoding, the next amino acid is delivered in a ternary complex of elongation factor Tu (EF-Tu), GTP and aminoacyl-tRNA. Decoding is followed by peptide-bond formation, resulting in the elongation of the polypeptide chain by one amino acid. EF-G-catalysed translocation moves the tRNAs and mRNA with respect to the ribosome.

Decoding. Decoding ensures that the correct aminoacyl-tRNA, as dictated by the mRNA codon, is selected in the A site. The binding of the appropriate ternary complex in the A site of the ribosome results in GTP hydrolysis by EF-Tu, the dissociation of the factor from the

ribosome and the movement of the aminoacyl end of A-site tRNA into the PTC, termed accommodation (Fig. 3). The many steps of decoding have been dissected by pre-steady state kinetic measurements²⁶ and single-molecule FRET studies²⁷.

The high accuracy of tRNA selection cannot be accounted for by just the free energy differences between base pairing and mismatches of the codon and anticodon^{28,29}, even considering the contribution of proofreading. Instead, interactions made by three universally conserved bases of the ribosome with the minor groove of the first two base pairs of the codon–anticodon helix gives rise to further discrimination (Fig. 3)³⁰. Such close monitoring of base-pairing geometry by the ribosome does not occur at the wobble position, consistent with the degeneracy of the genetic code. The binding energy of these extra interactions is not used primarily to increase the relative affinity of cognate versus near-cognate tRNA, but instead to induce a domain closure in the 30S subunit³¹, which presumably leads to the acceleration observed in rates of the forward steps in decoding³².

CryoEM studies of EF-Tu at increased resolution^{33,34} show that EF-Tu contacts the shoulder domain of the 30S subunit. Thus, domain closure would move the shoulder domain of the 30S subunit towards the ternary complex²⁹, potentially stabilizing the transition state for GTP hydrolysis by EF-Tu³¹ and leading to an acceleration of GTPase activation and tRNA selection. It seems that mutations or antibiotics that facilitate domain closure decrease the accuracy of the ribosome, whereas mutations that make domain closure more difficult result in increased accuracy^{29,31}.

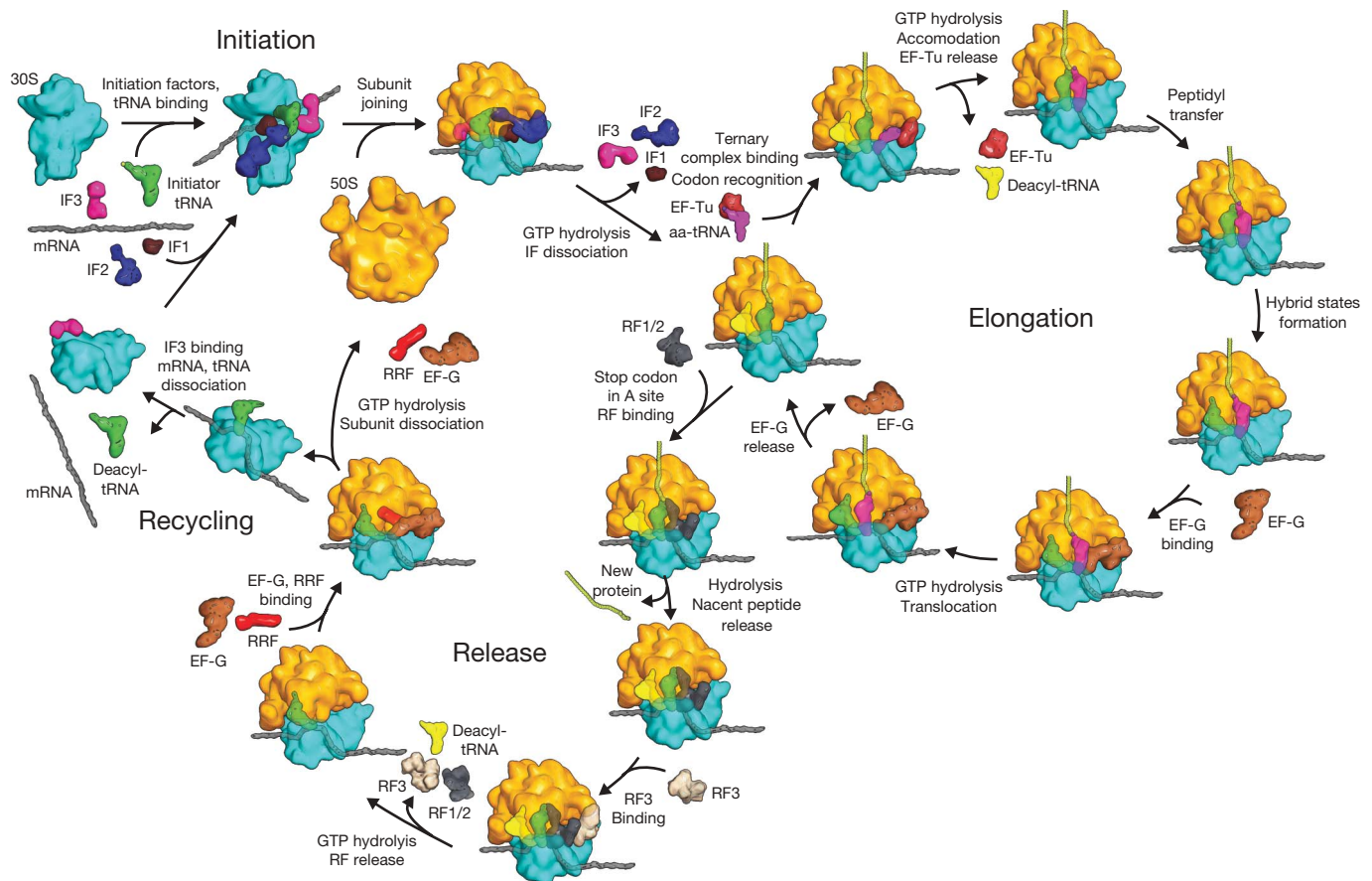


Figure 2 | Overview of bacterial translation. For simplicity, not all intermediate steps are shown. The colour scheme shown here is used consistently throughout this review. aa-tRNA, aminoacyl-tRNA; EF elongation factor; IF, initiation factor; RF, release factor.

These cryoEM structures, the most recent of which are beyond 7 Å resolution^{35,36}, also show that the tRNA is bent at the anticodon stem (Fig. 3f). The anticodon stem in the decoding centre is very nearly in the orientation acquired after accommodation and movement of the acceptor arm into the PTC. Thus, the binding energy derived from base pairing between the correct codon–anticodon is not only used to induce a conformational change in the ribosome, but also to distort the tRNA. A distorted tRNA may be characteristic of the transition state for GTP hydrolysis by EF-Tu, consistent with experiments

showing that a fragmented tRNA is unable to carry out decoding³⁷. In addition, recent mutational data on S12, a protein at the shoulder of the 30S subunit with a tail that stretches into the decoding centre, suggest it may be involved in relaying changes induced at the decoding centre to the ternary complex³⁸.

As this review was going to press, the crystal structure of EF-Tu and tRNA bound to the ribosome was determined³⁹. This structure shows details of the tRNA distortion that allows aminoacyl-tRNA to interact with both EF-Tu at the factor-binding site and the decoding centre of the 30S subunit. Furthermore, a series of conformational changes in aminoacyl-tRNA and EF-Tu that occur after productive ribosome binding suggest a communication pathway between the decoding centre and the GTPase centre of EF-Tu, which would trigger GTP hydrolysis after codon recognition.

After release of EF-Tu, the tRNA relaxes into the PTC^{31,34}. If the anticodon stem loop is held tightly at the decoding centre (as in the closed form induced by cognate tRNA), accommodation is accelerated⁴⁰. However, recent work on the Hirsh suppressor tRNA (a mutant Trp tRNA that recognizes the UGA stop codon) shows that this tRNA leads to acceleration of GTP hydrolysis and apparently accommodation with a near-cognate codon–anticodon pairing⁴¹. Thus, the mutant tRNA may be stabilized by additional interactions with the ribosome, rather than simply showing enhanced flexibility.

The discrimination achieved by monitoring the minor groove geometry in the codon–anticodon helix by decoding centre nucleotides though A-minor interactions can potentially yield an accuracy of $\sim 10^3$ – 10^4 in a single step⁴². Should the ribosome use this discrimination, then with proofreading, it would be possible to obtain much higher accuracy than is usually reported. Evidently, the ribosome forgoes accuracy by using the binding energy of codon–anticodon recognition to induce conformational changes in the ribosome and tRNA that result in accelerated GTP hydrolysis and tRNA selection.

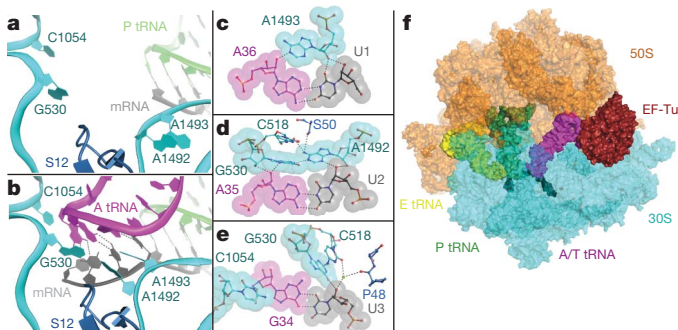


Figure 3 | Decoding by the ribosome. **a**, In the apo ribosome, A1492 and A1493 are stacked in h44. **b**, When a cognate tRNA binds to mRNA in the A site, A1492, A1493 and G530 change conformation to interact with the minor groove of the mRNA–tRNA mini-helix³⁰. **c–e**, Interactions of the 30S with the codon–anticodon pair. In the first (**c**) and second (**d**) positions, ribosomal bases monitor the geometry of the minor groove of the base pairs. Protein S12 also interacts with the second and third (**e**) positions. **f**, The ternary complex of EF-Tu and aminoacyl-tRNA with the 70S ribosome shows that the tRNA is bent in the anticodon stem (for example, see refs 35, 36).

However, a recent result suggests that the ribosome is capable of combining very high accuracy ($>10^6$) with a speed comparable to that of *in vivo* protein synthesis (~ 22 amino acids added per second)⁴³, both of which are much higher than previous measurements *in vitro* (accuracy ~ 450 , speed $\sim 6.6 \text{ s}^{-1}$)³². In the recent experiments⁴³, the accommodation of tRNA into the PTC is apparently too fast to allow significant discrimination by proofreading after GTP hydrolysis. If so, the structural basis of how one could have such a high accuracy with little or no proofreading is not clear, nor why measured *in vivo* rates of misincorporation are so much higher (reviewed in ref. 29). Further experiments with other reporters and in varying conditions are required to clarify these differences.

Peptide-bond formation. The central chemical event in protein synthesis is the peptidyl-transferase reaction, in which the α -amino group of the aminoacyl-tRNA nucleophilically attacks the ester carbon of the peptidyl-tRNA to form a new peptide bond (Fig. 4a; see the movie at http://www.sciencedirect.com/science/MiamiMultiMediaURL/B6WSR-4HHX2B2-B/B6WSR-4HHX2B2-B-2/7053/html/d074e3c1ecf8e4064d37dd72bc0b7e93/Movie_S1..mov). The ribosome increases the rate of this reaction by at least $\sim 10^5$ -fold⁴⁴. The catalytic site is in domain 5 of the 23S RNA, which binds the CCA ends of aminoacyl- and peptidyl-tRNA (Fig. 4b). It was located precisely in crystal structures of the *H. marismortui* 50S subunit⁴⁵, at the bottom of a large cleft (Fig. 4b). These structures precipitated many studies aimed at determining the catalytic mechanism of the peptidyl-transferase reaction. An initial proposal for a general acid/base catalytic mechanism involving N3 of A2451—a nucleotide in very close proximity to substrate analogues^{45,46}—was disproved by the dispensability of A2451 for the peptidyl-transferase reaction^{47–51}. Furthermore, crystal structures with improved resolution and more accurate transition state mimics showed that N3 of A2451 is not within hydrogen-bonding distance of the nucleophile throughout the reaction^{52,53}. When the reactive α -amine was substituted with a hydroxyl, making chemistry

rate limiting⁵⁴, a pH-independent reaction rate was observed. This is strong evidence that there is no general acid/base catalysis involving a group with near-neutral pK_a on A2451 or any other ribosomal moiety.

If there is no acid/base catalysis, what is the source of catalytic power of the ribosome? As with all enzymes, the precise organization of substrates and the active site plays an important contribution. In the ribosome, this is achieved when the binding of aminoacyl-tRNA induces a conformational change of the PTC and peptidyl-tRNA⁵⁵ (<http://www.nature.com/nature/journal/v438/n7067/extref/nature04152-s6.mov>, <http://www.nature.com/nature/journal/v438/n7067/extref/nature04152-s7.mov>). The α -amino group of the aminoacyl-tRNA interacts with the N3 of A2451 and the 2' OH of A76 of the peptidyl-tRNA, as part of an extensive network of hydrogen bonds that position the substrates for reaction (Fig. 4c)^{53,56–58}. It had been proposed that binding and orienting of substrates accounts for most of the ribosomal rate enhancement⁵⁹. A comparison of the rate of peptide-bond formation by the ribosome and by a ribosome-free model system suggested that the ribosome accelerated the reaction solely by entropic effects⁴⁴, which may include substrate positioning, shielding the reaction from bulk solvent, or organization of the active site^{44,57}. A precisely positioned water molecule interacts with the highly polarized transition state, as an oxyanion hole^{44,53}.

Although structural and biochemical studies have found no ribosomal group that acts in chemical catalysis, a substrate-assisted mechanism is possible. The 2' OH of the peptidyl-tRNA is well positioned to abstract and donate protons from the nucleophile and leaving group, respectively^{52,53,55}. Several studies suggest that this hydroxyl is vital for the reaction^{57,60–63}, whereas one group proposes it is dispensable⁶⁴. In the most rigorous study, Weinger *et al.*⁶³ substituted the 2' OH of A76 of peptidyl-tRNA with H or F (ref. 63), and found a rate reduction of at least 10^6 -fold. The importance and proximity of the 2' OH led to the proposal of a concerted proton shuttling mechanism, whereby it simultaneously accepts a proton from the α -amino group and donates one to the 3' O leaving group, perhaps as part of a six-membered ring of interactions^{53,57,62} (Fig. 4d, e). Such a mechanism may not require perturbation of the pK_a of the 2' OH pK_a .

Many mechanistic insights and biochemical experiments of the peptidyl-transferase reaction are based on structures of *H. marismortui* 50S complexes. It was questioned whether this reductionist system, which only includes the large subunit and the terminal nucleotides of tRNA, accurately represents the process in the whole ribosome with intact substrates^{7,65,66}. However, the 50S can catalyse the peptidyl-transferase reaction at similar rates to the 70S ribosome using a small dinucleotide A-site substrate, provided that a full-length tRNA is present in the P site⁶⁷. This analogue also shows the same robustness against active site mutations, and a pH profile similar to full aminoacyl-tRNAs in 70S ribosomes⁶⁸. Finally, recent structures show that a 70S ribosome with full-length tRNA substrates show that the PTC and substrate conformations are essentially identical to those in structures of the 50S with substrate analogues⁶⁹.

Although the ribosome is asymmetric, a pseudo-two-fold axis of symmetry exists at the PTC, relating the A and P sites⁶⁵. It is likely that 23S RNA started as a molecule of around 100 nucleotides, which duplicated to allow the proto-ribosome to bring two (non-coded) substrates into proximity^{65,70}. Careful analyses of the tertiary interactions reveal an evolutionary pathway of expansion of this proto-ribosome, giving rise to 23S RNA⁷⁰.

Recent studies shed light on the role of two proteins previously implicated in peptidyl transfer. In bacteria, the amino terminus of L27 could be crosslinked to the 3' end of both A- and P-site tRNAs, showing it was part of the PTC^{71,72}. Deletion or N-terminal truncation of L27 results in reduced peptidyl-transferase activity⁷² and computer simulations suggest the role of L27 is to aid binding of aminoacyl-tRNA⁷³. In addition, deletion of L16 was shown to cause a deficiency in A-site tRNA binding and the rate of peptidyl transfer^{74,75}. Recent structures show that the N-terminal tail of L27 is

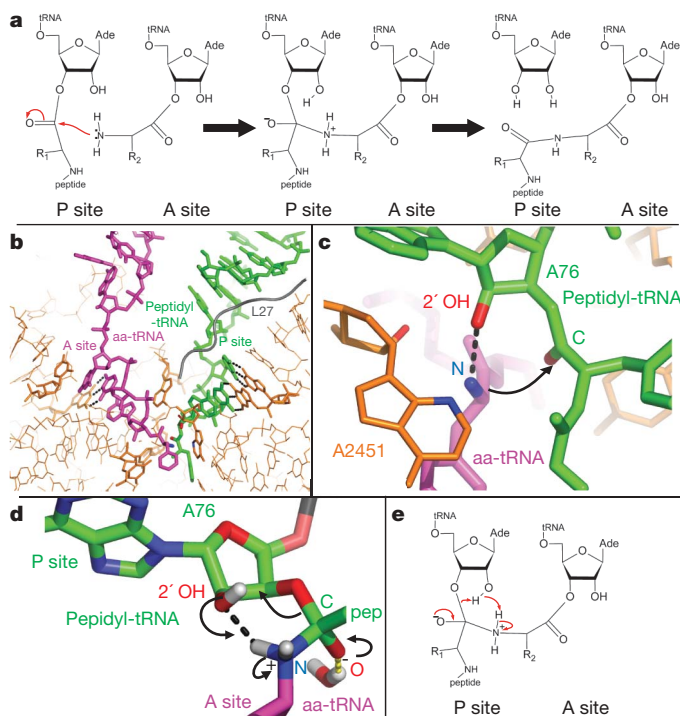


Figure 4 | Peptide-bond formation. **a**, Schematic drawing of the reaction. Ade, adenine. **b**, Binding of tRNAs to the PTC⁶⁹. **c**, The α -amino nucleophile is positioned by interaction with the 2' OH of A76 of peptidyl-tRNA and N3 of A2451, as part of an extensive network of hydrogen bonds⁵³. **d**, **e**, Possible mechanism by which the intermediate of the reaction breaks down into products, by a proton shuttle involving the 2' OH of A76 of peptidyl-tRNA.

ordered in the PTC where it interacts with the tRNA substrates^{8,69} (Fig. 4b), and that L16 becomes ordered owing to its interactions with the acceptor arm of A-site tRNA, rationalizing these findings. Thus, some proteins seem to aid the RNA components that primarily facilitate the peptidyl-transfer activity of the ribosome.

Translocation: the formation of hybrid states. With peptide-bond formation, the nascent peptide chain is transferred to the A-site tRNA leaving a deacylated tRNA in the P site. Before the next round of elongation, the tRNAs and mRNA need to move relative to the ribosome. During translocation the mRNA shifts by precisely one codon, except when either errors or programmed frameshifts occur. The tRNAs must also translocate from the A and P sites to the P and E sites, requiring a movement as large as 50 Å for the 3' end of the P-site tRNA.

Chemical footprinting showed that movements of the tRNAs occurred first with respect to the 50S subunit. P/E and A/P hybrid tRNA states form spontaneously after peptide-bond formation, and only after the addition of the GTPase elongation factor G (EF-G) did movement occur with respect to the 30S subunit⁷⁶ (Fig. 5). The hybrid states were visualized by careful sorting of ribosomal complexes^{77,78}. The ribosomal subunits have rotated by ~6° relative to each other⁷⁹ (<http://www.nature.com/nature/journal/v406/n6793/extref/406318ai1.mov>), and this 'ratcheted' ribosome contains both A/P and P/E tRNAs. Single-molecule FRET studies show that although the ribosome is initially in the unratcheted state, it oscillates between the unratcheted and ratcheted states after peptidyl transfer^{80,81}, until EF-G binding stabilizes the latter. To demonstrate that the ratcheted state of the ribosome is related to translocation, FRET experiments have shown that viomycin, an antibiotic that inhibits translocation, traps the ribosome in a ratcheted state indistinguishable by FRET from that obtained when EF-G is bound⁸². Furthermore, FRET measurements show that concomitantly with the ratcheting of the subunits, the L1 stalk moves to interact with the newly deacylated P-site tRNA, as would be expected if it moves into a P/E hybrid state⁸³. The formation of hybrid tRNA states is ordered, with the P/E tRNA state formed first, followed by the A/P state^{84,85}.

The role of the E site in translocation. The adoption of the P/E hybrid state also explains why the E site may be necessary. The E site is known to have evolved before the divergence of the three kingdoms, as the interactions of the E-site tRNA with the 50S subunit in archaea⁸⁶ and bacteria^{8,66} are similar. Because the E site binds deacylated but not peptidyl-tRNA, it is able to trap a hybrid P/E tRNA as soon as the P-site tRNA becomes deacylated, facilitating translocation by the formation of hybrid states. This concept is supported by

direct kinetic evidence⁸⁷ and the observation that tRNA modifications that affect E-site binding also affect translocation^{88,89}.

The crucial interactions between the ribosome and the terminal adenine of the E-site tRNA require only the 23S rRNA⁸⁶. Thus, the E site may have evolved before the evolution of proteins, such as translational factors that facilitate translocation by hybrid states. Consistent with this theory, ribosomes with modifications in S12 or S13 can perform translation even in the absence of elongation factors^{90,91}. These proteins are at the subunit interface, and their modification presumably disrupts contacts between the two subunits, facilitating their rotation relative to each other.

The role of EF-G in translocation relative to the 30S subunit. The second step in translocation is the movement of tRNAs and mRNA with respect to the 30S subunit, which is catalysed by EF-G. It is generally accepted that GTP hydrolysis by EF-G precedes translocation⁹².

CryoEM structures (for example, ref. 21) show that EF-G in the GTP-bound form on the ribosome has a significantly altered conformation from that in the GDP or apo form in isolation^{93,94}, and binds to the ratcheted state of the ribosome. A recent higher-resolution structure shows that the switch I and II regions of the GTPase domain become ordered on binding to the ribosome⁹⁵. Calorimetric studies suggest that EF-G undergoes a conformational change on binding GTP even before binding the ribosome, although full activation occurs only after ribosomal binding⁹⁶. The sarcin–ricin loop is the ribosomal element closest to the switch II region that is functionally important for GTPase activation in both EF-G and EF-Tu^{35,36,95}.

Ribosomes depleted of the L7/L12 stalk of the 50S subunit can bind EF-Tu or EF-G, but cannot efficiently activate GTP hydrolysis by the factors⁹⁷. L7 is an N-terminally modified form of L12, and by the association of its N-terminal domains exists as a tetramer in *Escherichia coli* or a hexamer in other species^{10,98}. This multimer of L12 binds a single copy of L10 to form a stalk that is fully or partially disordered in high-resolution structures of the ribosome. The tip of the stalk containing the C-terminal domain of L12 seems to be too far from the ribosomal GTPase centre to be involved directly in stimulating hydrolysis (Fig. 1). The structure of a hexamer of L12 complexed with L10 has been determined and modelled into the structure of the 50S subunit¹⁰. An increased rate of initial binding of GTPase factors was observed kinetically, which could be caused by several copies of the C-terminal domain of L12 effectively increasing the local concentration of the binding sites. Because the N- and C-terminal domains of L12 are connected by a flexible linker, the latter could move with the factor close to the sarcin–ricin loop.

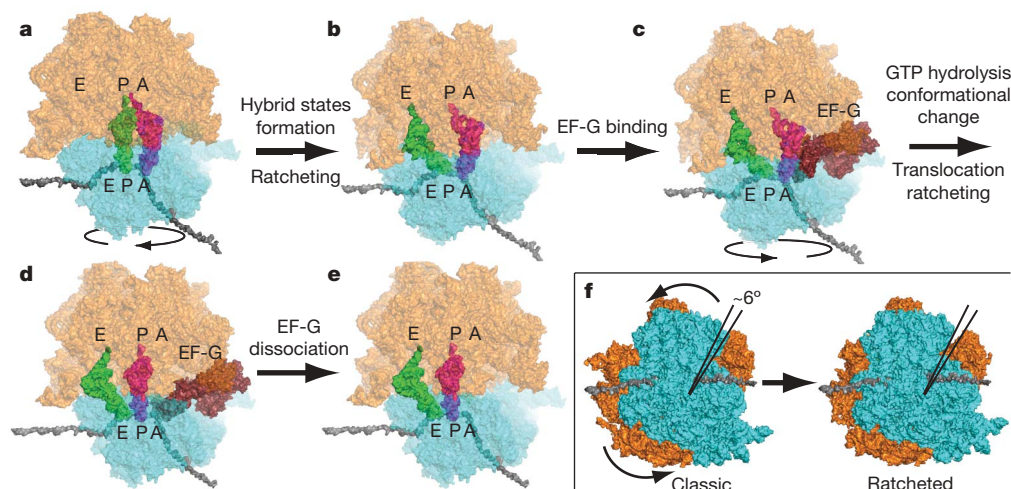


Figure 5 | EF-G catalysed translocation. a–e, After peptidyl transferase, tRNAs shift spontaneously to the A/P and P/E states in a ratcheted ribosome (b), to which EF-G binds. After GTP hydrolysis and tRNA movement, ratcheting reverses (d) and EF-G dissociates (e). f, Ratcheting involves a

rotation of the 30S subunit by approximately 6 degrees. See text for details. Note that transitions a-to-b, and c-to-d could be divided into sub-steps. No structure exists for c, so a domain movement was modelled to prevent EF-G and A/P tRNA from clashing.

The L11 region is in close proximity to the GTPase centre. This region also binds the antibiotics thiostrepton and micrococin, which inhibit and enhance GTPase activity, respectively. In the structure of the 50S subunit bound with micrococin, density for the C-terminal domain of L12 is observed adjacent to L11 and would be positioned to interact with EF-G⁹⁹. Thus micrococin could stabilize the binding of the C-terminal domain of L12 to the GTPase factor.

How GTP hydrolysis leads to movement of mRNA and tRNAs and resets the ratcheted ribosome to its canonical form is still unclear. Presumably, a rearrangement in the ribosome induced by GTP hydrolysis allows movement of the tRNAs and mRNA^{85,100}. Direct monitoring of mRNA showed that mRNA and tRNA movements occur at the same rate, and thus are directly coupled¹⁰¹. In the GDP form, the conformation of domain IV of EF-G places it in the A site of the 30S subunit^{102,103}. Thus, GTP hydrolysis may drive translocation by preventing the reverse movement of tRNA and mRNA. GTP hydrolysis may also allow the ribosome to act as a helicase and unwind the secondary structures formed in mRNA¹⁰⁴.

As this review was going to press, the crystal structure of EF-G•GDP trapped in the post-translocational state on the ribosome by fusidic acid was determined¹⁰⁵. The structure shows that domain IV makes extensive interactions with the minor groove of the codon–anticodon base pairs at the P site, but not with the A site codon. Fusidic acid seems to trap EF-G in a conformation between that of the GTP and GDP states, and the binding of EF-G in this state stabilizes the L10–L7/L12 stalk as well as (indirectly) the L1 stalk.

Termination of translation

The elongation cycle continues until an mRNA stop codon moves into the A site, signalling the end of the coding sequence. A class I release factor recognizes the stop codon and cleaves the nascent polypeptide chain from the P-site tRNA, resulting in the release of the newly synthesized protein from the ribosome. In bacteria, there are two class I release factors, RF1 and RF2. Whereas both factors recognize the UAA stop codon, UAG and UGA are only recognized by RF1 and RF2, respectively. In eukaryotes, a single eRF1 that is unrelated to RF1 or RF2 (refs 106, 107) recognizes all three stop codons. Tripeptide motifs PXT in RF1 and SPF in RF2 confer specificity for the codons UAG or UGA¹⁰⁸, whereas a universally conserved GGQ motif is implicated in peptide hydrolysis by release factors^{106,107}.

Unlike the extended structure of eRF1 (ref. 107), the crystal structure of isolated RF2 is compact, with the GGQ and SPF motifs only 23 Å apart¹⁰⁹. However, low-resolution structures showed that when bound to the ribosome, release factors were in an open form and domain 3, containing the GGQ motif, inserted into the PTC^{110–112}

(Fig. 6a). The high-resolution crystal form of the ribosome⁸ was used to solve structures of class I release factors bound to the bacterial ribosome^{113–115}, considerably advancing our understanding of the function of RF1 and RF2.

Recognition of the stop codon. Release-factor binding causes the conserved decoding centre bases G530, A1492 and A1493 to change conformation and form crucial interactions (Fig. 6b)^{113–115}. The changes are distinct from those during decoding of tRNA in which the bases monitor base-pairing geometry (Fig. 3c–e)³⁰, in a conformation incompatible with the binding of release factors¹¹⁶. Instead, A1493 stacks on A1913 of 23S RNA, forming a new contact between the two subunits, and G530 stacks onto the third stop codon base.

Although the structures can rationalize the specificity of RF1 and RF2 for their respective stop codons, the tripeptide motifs implicated in conferring specificity of RF1 or RF2 (ref. 108) make only limited interactions with the stop codon (Fig. 6b)^{113–115}, so that their role is still unclear. Bases after the stop codon affect release-factor efficiency (for example, ref. 117), and a single mutation in RF2 distant from the stop codon allows it to recognize all three stop codons¹¹⁸. Furthermore, it has been shown that a mismatch in the P site (after a near-cognate tRNA has been accepted and translocated) leads to release factor recognition of sense codons with increased efficiency¹¹⁹. Thus, release factor function may involve a subtle balance between the energetics of binding and conformational changes, similar to that during decoding by tRNA. After stop-codon recognition, peptide release is triggered, but the mechanism of signal transduction is unclear.

Catalysis of peptide release. The conserved GGQ motif is positioned in the PTC in a conformation only allowed because of its glycines (Fig. 6c)^{113–115}, explaining the drastic reduction in the activity of release factors when they are mutated^{120,121}. Furthermore, after release-factor binding U2585 shifts to expose the ester bond between the nascent peptide and P-site tRNA as is observed upon binding of A-site tRNA or its analogues^{55,69}. This shift has been proposed to be catalytically important for both peptidyl transfer and release⁵⁵, as in both cases it exposes the ester bond to attack by a nucleophile. Further supporting this theory, a variety of nucleophiles are effective to varying degrees during catalysis of peptide release by release factors¹²¹.

The glutamine in the GGQ motif makes a hydrogen bond through its main-chain amide with the 3' OH of A76 of deacylated P-site tRNA^{113–115}, which represents the product state after catalysis and release of the nascent peptide chain. One group has proposed product stabilization as part of the catalytic mechanism of release factors,

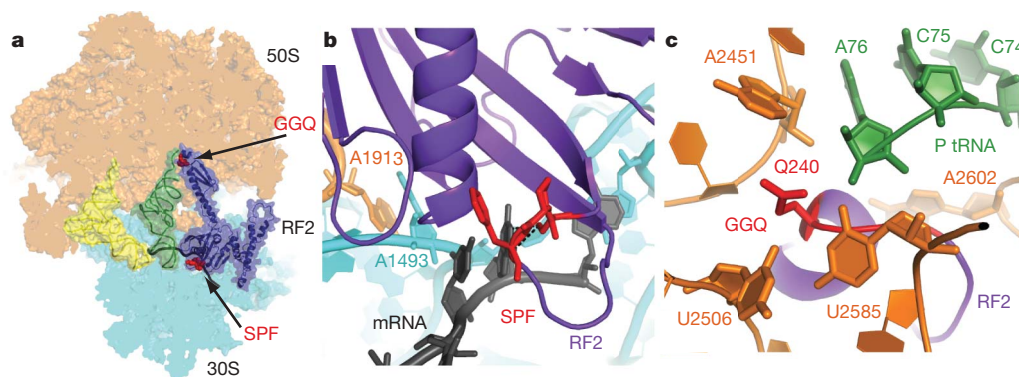


Figure 6 | Termination of translation by class I release factors. **a**, Overview of class I release-factor binding to the 70S ribosome^{113–115}. The view shows RF2; the GGQ motif implicated in catalysis of peptide release at the PTC and the SPF motif implicated in stop-codon recognition at the decoding centre

are highlighted in red. **b**, RF2 in the decoding centre, with the SPF motif highlighted. **c**, The PTC of the ribosome showing the GGQ motif of RF2 and deacylated P-site tRNA.

similar to certain proteases^{113,115}. Although the mutation of the glutamine to alanine results in only a modest 5–10-fold reduction in the catalytic rate^{120,121}, several other observations argue that it has a specific role in catalysis. The glutamine is universally conserved, which for glutamine is rarely for purely structural reasons. It is required for viability in bacteria and in eukaryotes^{107,122}. Whereas the mutation of the glutamine does not affect the rate of catalysis by other nucleophiles, it does specifically affect the rate for peptide release by water¹²¹. Finally, the side-chain amino group of the glutamine is methylated, and the loss of methylation was shown to reduce the efficiency of peptide release¹²³. Consistent with these data, one of the structural studies proposed a model in which the glutamine side chain directly coordinates a water molecule for nucleophilic attack¹¹⁴. Structures of the substrate and transition state complexes will address this question. **The role of RF3.** The class II release factor RF3 accelerates the dissociation of class I factors from the ribosome after peptide release. The binding of RF3 to the ribosome–RF1/2 complex in the GDP form is thought to induce RF3 to exchange GDP for GTP¹²⁴. The crystal structure of RF3–GDP resembles EF-Tu in the GTP form¹²⁵. The same study showed that the binding of RF3 in the GTP form to the ribosome induces conformational changes likely to destabilize the binding of class I release factors, thus leading to their dissociation from the ribosome.

Recycling of ribosomes before reinitiation. After hydrolysis of GTP by RF3, the factor dissociates from the ribosome, leaving mRNA and a deacylated tRNA in the P site. The ribosome must be recycled into subunits for a new round of protein synthesis to begin. In bacteria, an essential protein called ribosome recycling factor (RRF) works together with EF-G to carry out this process¹²⁶.

Chemical probing, cryoEM and crystallography all suggest similar interactions of RRF with the ribosome^{127–132}. However, the location of RRF in these studies would be incompatible with a P-site tRNA in the 50S subunit, as it would clash with the tip of domain I of RRF. Therefore, a probable model is that RRF binds to a ribosome containing a deacylated hybrid P/E tRNA. EF-G would then bind, similar to the way in which it binds the ribosome in a pre-translocation state. However, this view is complicated by studies suggesting that RRF can even act on ribosomes with a peptidyl-tRNA^{133,134}. CryoEM studies on the 50S subunit with RRF and both RRF and EF-G suggest the type of changes that might occur before and after RRF activity¹²⁹, but it is unclear if they represent any specific state of recycling. So far, there is no structure of the entire ribosome with both EF-G and RRF.

GTP hydrolysis seems to be required to promote the separation of subunits¹², yielding a 50S subunit and a complex of 30S, mRNA and deacylated tRNA, which requires IF3 to dissociate¹¹. The action of IF3 to remove mRNA and tRNA from the 30S subunit is attractive because it couples the last step in protein synthesis to the first, by preparing the 30S subunit for a new round of initiation (Fig. 2).

Conclusions

In this review, we have focused on the main aspects of bacterial translation that are common to the synthesis of all proteins. Although even this basic pathway is very complicated, translation involves many other features that have also been the subject of structural and functional studies in recent years. These include the rescue of stalled ribosomes, programmed frameshifting, the interaction of the nascent peptide with the exit tunnel, the modification of the peptide as it emerges from the ribosome, its folding and its transport across or insertion into membranes, and the regulation of translation. Nevertheless, one can only look back in wonder at the rate of progress in the last decade in our understanding of many key aspects of the translation pathway.

This progress is likely to continue unabated, with cryoEM now yielding increasingly high resolution and an increasing number of functional states becoming amenable to crystallographic studies. Two areas that would particularly benefit from high-resolution structures are initiation and translocation. These, as well as other stages of

translation involve GTPase factors. The very recent crystal structures of EF-Tu³⁹ and EF-G¹⁰⁵ represent the first high-resolution structures of GTPase factors bound to the ribosome. By showing that such complexes are accessible to crystallography, they allow us to be optimistic that similar structures in other states will lead to an understanding of how the ribosome specifically activates GTP hydrolysis by these factors at precise stages that differ for each GTPase. In addition to structural studies, increasingly sophisticated biochemical methods such as single-molecule studies will help to dissect the various steps of complicated processes. Although the structures of key states of the ribosome will be welcome, an understanding of how the ribosome proceeds from one state to the next will be aided by molecular dynamics, which is now able to tackle larger and more complex problems as a result of advances in computing and methodology. Finally, the extremely complicated field of eukaryotic translation, especially initiation, is sure to be increasingly targeted by biophysical and biochemical techniques.

Published online 18 October 2009.

- Ramakrishnan, V. Ribosome structure and the mechanism of translation. *Cell* **108**, 557–572 (2002).
- Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
- Wimberly, B. T. et al. Structure of the 30S ribosomal subunit. *Nature* **407**, 327–339 (2000).
- Yusupov, M. M. et al. Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292**, 883–896 (2001).
- Gao, H. et al. Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. *Cell* **113**, 789–801 (2003).
- Harms, J. et al. High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107**, 679–688 (2001).
- Schuwirth, B. S. et al. Structures of the bacterial ribosome at 3.5 Å resolution. *Science* **310**, 827–834 (2005).
- Selmer, M. et al. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science* **313**, 1935–1942 (2006).
This high-resolution structure of a functional complex of the ribosome has paved the way for many other studies.
- Nikulin, A. et al. Structure of the L1 protuberance in the ribosome. *Nature Struct. Biol.* **10**, 104–108 (2003).
- Diaconu, M. et al. Structural basis for the function of the ribosomal L7/L12 stalk in factor binding and GTPase activation. *Cell* **121**, 991–1004 (2005).
- Karimi, R., Pavlov, M. Y., Buckingham, R. H. & Ehrenberg, M. Novel roles for classical factors at the interface between translation termination and initiation. *Mol. Cell* **3**, 601–609 (1999).
- Peske, F., Rodnina, M. V. & Wintermeyer, W. Sequence of steps in ribosome recycling as defined by kinetic analysis. *Mol. Cell* **18**, 403–412 (2005).
- Antoun, A., Pavlov, M. Y., Lovmar, M. & Ehrenberg, M. How initiation factors maximize the accuracy of tRNA selection in initiation of bacterial protein synthesis. *Mol. Cell* **23**, 183–193 (2006).
- Grigoriadou, C., Marzi, S., Pan, D., Gualerzi, C. O. & Cooperman, B. S. The translational fidelity function of IF3 during transition from the 30 S initiation complex to the 70 S initiation complex. *J. Mol. Biol.* **373**, 551–561 (2007).
- Milon, P., Konevega, A. L., Gualerzi, C. O. & Rodnina, M. V. Kinetic checkpoint at a late step in translation initiation. *Mol. Cell* **30**, 712–720 (2008).
- Tomsic, J. et al. Late events of translation initiation in bacteria: a kinetic analysis. *EMBO J.* **19**, 2127–2136 (2000).
- Grigoriadou, C., Marzi, S., Kirillov, S., Gualerzi, C. O. & Cooperman, B. S. A quantitative kinetic scheme for 70 S translation initiation complex formation. *J. Mol. Biol.* **373**, 562–572 (2007).
- Carter, A. P. et al. Crystal structure of an initiation factor bound to the 30S ribosomal subunit. *Science* **291**, 498–501 (2001).
- Simonetti, A. et al. Structure of the 30S translation initiation complex. *Nature* **455**, 416–420 (2008).
Refs 19 and 20 have shed light on the location of initiation factors in the ribosome.
- Allen, G. S., Zavialov, A., Gursky, R., Ehrenberg, M. & Frank, J. The cryo-EM structure of a translation initiation complex from *Escherichia coli*. *Cell* **121**, 703–712 (2005).
- Valle, M. et al. Locking and unlocking of ribosomal motions. *Cell* **114**, 123–134 (2003).
- Myasnikov, A. G. et al. Conformational transition of initiation factor 2 from the GTP- to GDP-bound state visualized on the ribosome. *Nature Struct. Mol. Biol.* **12**, 1145–1149 (2005).
- Allen, G. S. & Frank, J. Structural insights on the translation initiation complex: ghosts of a universal initiation complex. *Mol. Microbiol.* **63**, 941–950 (2007).
- Marshall, R. A., Aitken, C. E. & Puglisi, J. D. GTP hydrolysis by IF2 guides progression of the ribosome into elongation. *Mol. Cell* **35**, 37–47 (2009).

25. Antoun, A., Pavlov, M. Y., Andersson, K., Tenson, T. & Ehrenberg, M. The roles of initiation factor 2 and guanosine triphosphate in initiation of protein synthesis. *EMBO J.* **22**, 5593–5601 (2003).
26. Rodnina, M. V. & Wintermeyer, W. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu. Rev. Biochem.* **70**, 415–435 (2001).
27. Blanchard, S. C., Gonzalez, R. L., Kim, H. D., Chu, S. & Puglisi, J. D. tRNA selection and kinetic proofreading in translation. *Nature Struct. Mol. Biol.* **11**, 1008–1014 (2004).
28. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
29. Ogle, J. M. & Ramakrishnan, V. Structural insights into translational fidelity. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
30. Ogle, J. M. *et al.* Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science* **292**, 897–902 (2001).
31. Ogle, J. M., Murphy, F. V., Tarry, M. J. & Ramakrishnan, V. Selection of tRNA by the ribosome requires a transition from an open to a closed form. *Cell* **111**, 721–732 (2002).
32. Gromadski, K. B. & Rodnina, M. V. Kinetic determinants of high-fidelity tRNA discrimination on the ribosome. *Mol. Cell* **13**, 191–200 (2004).
33. Stark, H. *et al.* Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. *Nature Struct. Biol.* **9**, 849–854 (2002).
34. Valle, M. *et al.* Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. *Nature Struct. Biol.* **10**, 899–906 (2003).
35. Villa, E. *et al.* Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. *Proc. Natl Acad. Sci. USA* **106**, 1063–1068 (2009). **This paper and ref. 36 report cryoEM structures of an important state of the ribosome at and beyond 7 Å resolution.**
36. Schuetz, J. C. *et al.* GTPase activation of elongation factor EF-Tu by the ribosome during decoding. *EMBO J.* **28**, 755–765 (2009).
37. Piepenburg, O. *et al.* Intact aminoacyl-tRNA is required to trigger GTP hydrolysis by elongation factor Tu on the ribosome. *Biochemistry* **39**, 1734–1738 (2000).
38. Gregory, S. T., Carr, J. F. & Dahlberg, A. E. A signal relay between ribosomal protein S12 and elongation factor EF-Tu during decoding of mRNA. *RNA* **15**, 208–214 (2009).
39. Schmeing, T. M. *et al.* The structure of the ribosome bound to EF-Tu and tRNA. *Science* doi:10.1126/science.1179700 (in press).
40. Pape, T., Wintermeyer, W. & Rodnina, M. Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. *EMBO J.* **18**, 3800–3807 (1999).
41. Cochella, L. & Green, R. An active role for tRNA beyond codon:anticodon base pairing. *Science* **308**, 1178–1180 (2005).
42. Battle, D. J. & Doudna, J. A. Specificity of RNA-RNA helix recognition. *Proc. Natl Acad. Sci. USA* **99**, 11676–11681 (2002).
43. Johansson, M., Bouakaz, E., Lovmar, M. & Ehrenberg, M. The kinetics of ribosomal peptidyl transfer revisited. *Mol. Cell* **30**, 589–598 (2008).
44. Sievers, A., Beringer, M., Rodnina, M. V. & Wolfenden, R. The ribosome as an entropy trap. *Proc. Natl Acad. Sci. USA* **101**, 7897–7901 (2004).
45. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
46. Muth, G. W., Ortoleva-Donnelly, L. & Strobel, S. A. A single adenosine with a neutral pK_a in the ribosomal peptidyl transferase center. *Science* **289**, 947–950 (2000).
47. Beringer, M., Adio, S., Wintermeyer, W. & Rodnina, M. The G2447A mutation does not affect ionization of a ribosomal group taking part in peptide bond formation. *RNA* **9**, 919–922 (2003).
48. Polacek, N., Gaynor, M., Yassin, A. & Mankin, A. S. Ribosomal peptidyl transferase can withstand mutations at the putative catalytic nucleotide. *Nature* **411**, 498–501 (2001).
49. Thompson, J. *et al.* Analysis of mutations at residues A2451 and G2447 of 23S rRNA in the peptidyltransferase active site of the 50S ribosomal subunit. *Proc. Natl Acad. Sci. USA* **98**, 9002–9007 (2001).
50. Katunin, V. I., Muth, G. W., Strobel, S. A., Wintermeyer, W. & Rodnina, M. V. Important contribution to catalysis of peptide bond formation by a single ionizing group within the ribosome. *Mol. Cell* **10**, 339–346 (2002).
51. Youngman, E. M., Brunelle, J. L., Kochaniak, A. B. & Green, R. The active site of the ribosome is composed of two layers of conserved nucleotides with distinct roles in peptide bond formation and peptide release. *Cell* **117**, 589–599 (2004).
52. Hansen, J. L., Schmeing, T. M., Moore, P. B. & Steitz, T. A. Structural insights into peptide bond formation. *Proc. Natl Acad. Sci. USA* **99**, 11670–11675 (2002).
53. Schmeing, T. M., Huang, K. S., Kitchen, D. E., Strobel, S. A. & Steitz, T. A. Structural insights into the roles of water and the 2' hydroxyl of the P site tRNA in the peptidyl transferase reaction. *Mol. Cell* **20**, 437–448 (2005).
54. Bieling, P., Beringer, M., Adio, S. & Rodnina, M. V. Peptide bond formation does not involve acid-base catalysis by ribosomal residues. *Nature Struct. Mol. Biol.* **13**, 423–428 (2006).
55. Schmeing, T. M., Huang, K. S., Strobel, S. A. & Steitz, T. A. An induced-fit mechanism to promote peptide bond formation and exclude hydrolysis of peptidyl-tRNA. *Nature* **438**, 520–524 (2005). **This paper shows that peptidyl-transferase activity involves an induced conformational change that opens the ester bond between the peptide and tRNA to nucleophilic attack.**
56. Lang, K., Erlacher, M., Wilson, D. N., Micura, R. & Polacek, N. The role of 23S ribosomal RNA residue A2451 in peptide bond synthesis revealed by atomic mutagenesis. *Chem. Biol.* **15**, 485–492 (2008).
57. Trobro, S. & Aqvist, J. Mechanism of peptide bond synthesis on the ribosome. *Proc. Natl Acad. Sci. USA* **102**, 12395–12400 (2005).
58. Sharma, P. K., Xiang, Y., Kato, M. & Warshel, A. What are the roles of substrate-assisted catalysis and proximity effects in peptide bond formation by the ribosome? *Biochemistry* **44**, 11307–11314 (2005).
59. Nierhaus, K. H., Schulze, H. & Cooperman, B. S. Molecular mechanisms of the ribosomal peptidyltransferase center. *Biochem. Int.* **1**, 185–192 (1980).
60. Hecht, S. M., Kozarich, J. W. & Schmidt, F. J. Isomeric phenylalanyl-tRNAs. Position of the aminoacyl moiety during protein biosynthesis. *Proc. Natl Acad. Sci. USA* **71**, 4317–4321 (1974).
61. Quiggle, K., Kumar, G., Ott, T. W., Ryu, E. K. & Chladek, S. Donor site of ribosomal peptidyltransferase: investigation of substrate specificity using 2'(3')-O-(N-acylaminoacyl)dinucleoside phosphates as models of the 3' terminus of N-acylaminoacyl transfer ribonucleic acid. *Biochemistry* **20**, 3480–3485 (1981).
62. Dorner, S., Panuschka, C., Schmid, W. & Barta, A. Mononucleotide derivatives as ribosomal P-site substrates reveal an important contribution of the 2'-OH to activity. *Nucleic Acids Res.* **31**, 6536–6542 (2003).
63. Weinger, J. S., Parnell, K. M., Dorner, S., Green, R. & Strobel, S. A. Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nature Struct. Mol. Biol.* **11**, 1101–1106 (2004).
64. Koch, M., Huang, Y. & Sprinzl, M. Peptide-bond synthesis on the ribosome: no free vicinal hydroxy group required on the terminal ribose residue of peptidyl-tRNA. *Angew. Chem. Int. Edn Engl.* **47**, 7242–7245 (2008).
65. Bashan, A. *et al.* Structural basis of the ribosomal machinery for peptide bond formation, translocation, and nascent chain progression. *Mol. Cell* **11**, 91–102 (2003).
66. Korostelev, A., Trakhanov, S., Laurberg, M. & Noller, H. F. Crystal structure of a 70S ribosome-tRNA complex reveals functional interactions and rearrangements. *Cell* **126**, 1065–1077 (2006).
67. Wohlgemuth, I., Beringer, M. & Rodnina, M. V. Rapid peptide bond formation on isolated 50S ribosomal subunits. *EMBO Rep.* **7**, 699–703 (2006).
68. Brunelle, J. L., Youngman, E. M., Sharma, D. & Green, R. The interaction between C75 of tRNA and the A loop of the ribosome stimulates peptidyl transferase activity. *RNA* **12**, 33–39 (2006).
69. Voorhees, R. M., Weixlbaumer, A., Loakes, D., Kelley, A. C. & Ramakrishnan, V. Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome. *Nature Struct. Mol. Biol.* **16**, 528–533 (2009).
70. Bokov, K. & Steinberg, S. V. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* **457**, 977–980 (2009).
71. Wower, J., Hixson, S. S. & Zimmermann, R. A. Labeling the peptidyltransferase center of the *Escherichia coli* ribosome with photoreactive tRNA(Phe) derivatives containing azidoadenosine at the 3' end of the acceptor arm: a model of the tRNA-ribosome complex. *Proc. Natl Acad. Sci. USA* **86**, 5232–5236 (1989).
72. Maguire, B. A., Benjaminov, A. D., Ramu, H., Mankin, A. S. & Zimmermann, R. A. A protein component at the heart of an RNA machine: the importance of protein I27 for the function of the bacterial ribosome. *Mol. Cell* **20**, 427–435 (2005).
73. Trobro, S. & Aqvist, J. Role of ribosomal protein L27 in peptidyl transfer. *Biochemistry* **47**, 4898–4906 (2008).
74. Moore, V. G., Atchison, R. E., Thomas, G., Moran, M. & Noller, H. F. Identification of a ribosomal protein essential for peptidyl transferase activity. *Proc. Natl Acad. Sci. USA* **72**, 844–848 (1975).
75. Kazemie, M. Binding of aminoacyl-tRNA to reconstituted subparticles of *Escherichia coli* large ribosomal subunits. *Eur. J. Biochem.* **67**, 373–378 (1976).
76. Moazed, D. & Noller, H. F. Intermediate states in the movement of transfer RNA in the ribosome. *Nature* **342**, 142–148 (1989).
77. Agirreazabal, X. *et al.* Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Mol. Cell* **32**, 190–197 (2008). **This paper and ref. 78 show direct structural evidence for hybrid states following peptide-bond formation.**
78. Julián, P. *et al.* Structure of ratcheted ribosomes with tRNAs in hybrid states. *Proc. Natl Acad. Sci. USA* **105**, 16924–16927 (2008).
79. Frank, J. & Agrawal, R. K. A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature* **406**, 318–322 (2000).
80. Blanchard, S. C., Kim, H. D., Gonzalez, R. L. Jr, Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101**, 12893–12898 (2004). **A ground-breaking paper on the use of single molecule techniques to study dynamics in the ribosome.**
81. Cornish, P. V., Ermolenko, D. N., Noller, H. F. & Ha, T. Spontaneous intersubunit rotation in single ribosomes. *Mol. Cell* **30**, 578–588 (2008).
82. Ermolenko, D. N. *et al.* The antibiotic viomycin traps the ribosome in an intermediate state of translocation. *Nature Struct. Mol. Biol.* **14**, 493–497 (2007).
83. Fei, J., Kosuri, P., MacDougall, D. D. & Gonzalez, R. L. Jr. Coupling of ribosomal L1 stalk and tRNA dynamics during translation elongation. *Mol. Cell* **30**, 348–359 (2008).
84. Munro, J. B., Altman, R. B., O'Connor, N. & Blanchard, S. C. Identification of two distinct hybrid state intermediates on the ribosome. *Mol. Cell* **25**, 505–517 (2007).
85. Pan, D., Kirillov, S. V. & Cooperman, B. S. Kinetically competent intermediates in the translocation step of protein synthesis. *Mol. Cell* **25**, 519–529 (2007).

86. Schmeing, T. M., Moore, P. B. & Steitz, T. A. Structures of deacylated tRNA mimics bound to the E site of the large ribosomal subunit. *RNA* **9**, 1345–1352 (2003).
87. Savelsbergh, A., Mohr, D., Wilden, B., Wintermeyer, W. & Rodnina, M. V. Stimulation of the GTPase activity of translation elongation factor G by ribosomal protein L7/12. *J. Biol. Chem.* **275**, 890–894 (2000).
88. Lill, R., Robertson, J. M. & Wintermeyer, W. Binding of the 3' terminus of tRNA to 23S rRNA in the ribosomal exit site actively promotes translocation. *EMBO J.* **8**, 3933–3938 (1989).
89. Feinberg, J. S. & Joseph, S. Identification of molecular interactions between P-site tRNA and the ribosome essential for translocation. *Proc. Natl Acad. Sci. USA* **98**, 11120–11125 (2001).
90. Gavrilova, L. P., Kotliansky, V. E. & Spirin, A. S. Ribosomal protein S12 and 'non-enzymatic' translocation. *FEBS Lett.* **45**, 324–328 (1974).
91. Cukras, A. R., Southworth, D. R., Brunelle, J. L., Culver, G. M. & Green, R. Ribosomal proteins S12 and S13 function as control elements for translocation of the mRNA:tRNA complex. *Mol. Cell* **12**, 321–328 (2003).
92. Rodnina, M. V., Savelsbergh, A., Katunin, V. I. & Wintermeyer, W. Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome. *Nature* **385**, 37–41 (1997).
93. Evarsson, A. *et al.* Three-dimensional structure of the ribosomal translocase: elongation factor G from *Thermus thermophilus*. *EMBO J.* **13**, 3669–3677 (1994).
94. Czworkowski, J., Wang, J., Steitz, T. A. & Moore, P. B. The crystal structure of elongation factor G complexed with GDP, at 2.7 Å resolution. *EMBO J.* **13**, 3661–3668 (1994).
95. Connell, S. R. *et al.* Structural basis for interaction of the ribosome with the switch regions of GTP-bound elongation factors. *Mol. Cell* **25**, 751–764 (2007).
96. Haurlyuk, V. *et al.* The pretranslocation ribosome is targeted by GTP-bound EF-G in partially activated form. *Proc. Natl Acad. Sci. USA* **105**, 15678–15683 (2008).
97. Mohr, D., Wintermeyer, W. & Rodnina, M. V. GTPase activation of elongation factors Tu and G on the ribosome. *Biochemistry* **41**, 12520–12528 (2002).
98. Ilag, L. L. *et al.* Heptameric (L12)₆/L10 rather than canonical pentameric complexes are found by tandem MS of intact ribosomes from thermophilic bacteria. *Proc. Natl Acad. Sci. USA* **102**, 8192–8197 (2005).
99. Harms, J. M. *et al.* Translational regulation via L11: molecular switches on the ribosome turned on and off by thiostrepton and micrococin. *Mol. Cell* **30**, 26–38 (2008).
100. Savelsbergh, A. *et al.* An elongation factor G-induced ribosome rearrangement precedes tRNA-mRNA translocation. *Mol. Cell* **11**, 1517–1523 (2003).
101. Studer, S. M., Feinberg, J. S. & Joseph, S. Rapid kinetic analysis of EF-G-dependent mRNA translocation in the ribosome. *J. Mol. Biol.* **327**, 369–381 (2003).
102. Agrawal, R. K., Penczek, P., Grassucci, R. A. & Frank, J. Visualization of elongation factor G on the *Escherichia coli* 70S ribosome: the mechanism of translocation. *Proc. Natl Acad. Sci. USA* **95**, 6134–6138 (1998).
103. Stark, H., Rodnina, M. V., Wieden, H. J., van Heel, M. & Wintermeyer, W. Large-scale movement of elongation factor G and extensive conformational change of the ribosome during translocation. *Cell* **100**, 301–309 (2000).
104. Takyar, S., Hickerson, R. P. & Noller, H. F. mRNA helicase activity of the ribosome. *Cell* **120**, 49–58 (2005).
105. Gao, Y.-G. *et al.* The structure of the ribosome with elongation factor G trapped in the post-translocational state. *Science* doi:10.1126/science.1179709 (in the press).
106. Frolova, L. Y. *et al.* Mutations in the highly conserved GGQ motif of class 1 polypeptide release factors abolish ability of human eRF1 to trigger peptidyl-tRNA hydrolysis. *RNA* **5**, 1014–1020 (1999).
107. Song, H. *et al.* The crystal structure of human eukaryotic release factor eRF1-mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. *Cell* **100**, 311–321 (2000).
108. Ito, K., Uno, M. & Nakamura, Y. A tripeptide 'anticodon' deciphers stop codons in messenger RNA. *Nature* **403**, 680–684 (2000).
109. Vestergaard, B. *et al.* Bacterial polypeptide release factor RF2 is structurally distinct from eukaryotic eRF1. *Mol. Cell*, (2001).
110. Klaholz, B. P. *et al.* Structure of the *Escherichia coli* ribosomal termination complex with release factor 2. *Nature* **421**, 90–94 (2003).
111. Rawat, U. B. *et al.* A cryo-electron microscopic study of ribosome-bound termination factor RF2. *Nature* **421**, 87–90 (2003).
112. Petry, S. *et al.* Crystal structures of the ribosome in complex with release factors RF1 and RF2 bound to a cognate stop codon. *Cell* **123**, 1255–1266 (2005).
113. Laurberg, M. *et al.* Structural basis for translation termination on the 70S ribosome. *Nature* **454**, 852–857 (2008).
- This paper and ref. 114 provide insights into the recognition of stop codons by release factors.
114. Weixlbaumer, A. *et al.* Insights into translational termination from the structure of RF2 bound to the ribosome. *Science* **322**, 953–956 (2008).
115. Korostelev, A. *et al.* Crystal structure of a translation termination complex formed with release factor RF2. *Proc. Natl Acad. Sci. USA* **105**, 19684–19689 (2008).
116. Youngman, E. M., He, S. L., Nikstad, L. J. & Green, R. Stop codon recognition by release factors induces structural rearrangement of the ribosomal decoding center that is productive for peptide release. *Mol. Cell* **28**, 533–543 (2007).
117. Poole, E. S., Major, L. L., Mannering, S. A. & Tate, W. P. Translational termination in *Escherichia coli*: three bases following the stop codon crosslink to release factor 2 and affect the decoding efficiency of UGA-containing signals. *Nucleic Acids Res.* **26**, 954–960 (1998).
118. Ito, K., Uno, M. & Nakamura, Y. Single amino acid substitution in prokaryote polypeptide release factor 2 permits it to terminate translation at all three stop codons. *Proc. Natl Acad. Sci. USA* **95**, 8165–8169 (1998).
119. Zaher, H. S. & Green, R. Quality control by the ribosome following peptide bond formation. *Nature* **457**, 161–166 (2009).
120. Zavialov, A. V., Mora, L., Buckingham, R. H. & Ehrenberg, M. Release of peptide promoted by the GGQ motif of class 1 release factors regulates the GTPase activity of RF3. *Mol. Cell* **10**, 789–798 (2002).
121. Shaw, J. J. & Green, R. Two distinct components of release factor function uncovered by nucleophile partitioning analysis. *Mol. Cell* **28**, 458–467 (2007).
122. Mora, L. *et al.* The essential role of the invariant GGQ motif in the function and stability *in vivo* of bacterial release factors RF1 and RF2. *Mol. Microbiol.* **47**, 267–275 (2003).
123. Dinçbas-Renqvist, V. *et al.* A post-translational modification in the GGQ motif of RF2 from *Escherichia coli* stimulates termination of translation. *EMBO J.* **19**, 6900–6907 (2000).
124. Zavialov, A. V., Buckingham, R. H. & Ehrenberg, M. A posttermination ribosomal complex is the guanine nucleotide exchange factor for peptide release factor rf3. *Cell* **107**, 115–124 (2001).
125. Gao, H. *et al.* RF3 induces ribosomal conformational changes responsible for dissociation of class I release factors. *Cell* **129**, 929–941 (2007).
126. Hirashima, A. & Kaji, A. Role of elongation factor G and a protein factor on the release of ribosomes from messenger ribonucleic acid. *J. Biol. Chem.* **248**, 7580–7587 (1973).
127. Lancaster, L., Kiel, M. C., Kaji, A. & Noller, H. F. Orientation of ribosome recycling factor in the ribosome from directed hydroxyl radical probing. *Cell* **111**, 129–140 (2002).
128. Agrawal, R. K. *et al.* Visualization of ribosome-recycling factor on the *Escherichia coli* 70S ribosome: functional implications. *Proc. Natl Acad. Sci. USA* **101**, 8900–8905 (2004).
129. Gao, N. *et al.* Mechanism for the disassembly of the posttermination complex inferred from cryo-EM studies. *Mol. Cell* **18**, 663–674 (2005).
130. Wilson, D. N. *et al.* X-ray crystallography study on ribosome recycling: the mechanism of binding and action of RRF on the 50S ribosomal subunit. *EMBO J.* **24**, 251–260 (2005).
131. Borovinskaya, M. A. *et al.* Structural basis for aminoglycoside inhibition of bacterial ribosome recycling. *Nature Struct. Mol. Biol.* **14**, 727–732 (2007).
132. Weixlbaumer, A. *et al.* Crystal structure of the ribosome recycling factor bound to the ribosome. *Nature Struct. Mol. Biol.* **14**, 733–737 (2007).
133. Heurgué-Hamard, V. *et al.* Ribosome release factor RF4 and termination factor RF3 are involved in dissociation of peptidyl-tRNA from the ribosome. *EMBO J.* **17**, 808–816 (1998).
134. Rao, A. R. & Varshney, U. Specific interaction between the ribosome recycling factor and the elongation factor G from *Mycobacterium tuberculosis* mediates peptidyl-tRNA release and ribosome recycling in *Escherichia coli*. *EMBO J.* **20**, 2977–2986 (2001).

Acknowledgements We thank R. Voorhees for a critical reading of this manuscript, and J. Frank and X. Aggirrezabala for providing coordinates of a hybrid state complex. Work in V.R.'s laboratory is supported by the Medical Research Council (UK), the Wellcome Trust, the Louis-Jeantet Foundation and the Agouron Institute. T.M.S. was supported by fellowships from the Human Frontiers Science Program and Emmanuel College, Cambridge. Part of this review was written when V.R. was a G. N. Ramachandran Visiting Professor at the Indian Institute of Science, Bangalore, where he thanks U. Varshney for his hospitality and useful discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to V.R. (ramak@mrclmb.cam.ac.uk) or T.M.S. (schmeing@mrclmb.cam.ac.uk).

Genome evolution and adaptation in a long-term experiment with *Escherichia coli*

Jeffrey E. Barrick^{1*}, Dong Su Yu^{2,3*}, Sung Ho Yoon², Haeyoung Jeong², Tae Kwang Oh^{2,4}, Dominique Schneider⁵, Richard E. Lenski¹ & Jihyun F. Kim^{2,6}

The relationship between rates of genomic evolution and organismal adaptation remains uncertain, despite considerable interest. The feasibility of obtaining genome sequences from experimentally evolving populations offers the opportunity to investigate this relationship with new precision. Here we sequence genomes sampled through 40,000 generations from a laboratory population of *Escherichia coli*. Although adaptation decelerated sharply, genomic evolution was nearly constant for 20,000 generations. Such clock-like regularity is usually viewed as the signature of neutral evolution, but several lines of evidence indicate that almost all of these mutations were beneficial. This same population later evolved an elevated mutation rate and accumulated hundreds of additional mutations dominated by a neutral signature. Thus, the coupling between genomic and adaptive evolution is complex and can be counterintuitive even in a constant environment. In particular, beneficial substitutions were surprisingly uniform over time, whereas neutral substitutions were highly variable.

Adaptation has often been viewed as a gradual process. Darwin¹ wrote that “We see nothing of these slow changes in progress, until the hand of time has marked the long lapse of ages...”. Theoretical work in quantitative genetics supported this view by showing that gradual adaptation would result from constant selection on many mutations of small effect². However, an alternative model of evolution on rugged fitness landscapes challenged this perspective³ and, later, empirical evidence was found for alternating periods of rapid phenotypic evolution and stasis in some lineages^{4,5}. The causes of variation in the rate of adaptation remain controversial and are probably diverse. They may include changes in the environment, in circumstances promoting or impeding gene flow, and in opportunities for refinement following the origin of key innovations or the invasion of new habitats, among other factors^{6–11}.

Genomic changes underlie evolutionary adaptation, but mutations—even those substituted (fixed) in evolving populations—are not necessarily beneficial. Variation in the rate of genomic evolution is also subject to many influences and complications. On the one hand, theory predicts that neutral mutations should accumulate by drift at a uniform rate, albeit stochastically, provided the mutation rate is constant¹². On the other hand, rates of substitution of beneficial and deleterious mutations depend on selection, and hence the environment, as well as on population size and structure^{13,14}. Moreover, the relative proportions of substitutions that are neutral, deleterious and beneficial are usually difficult to infer given imperfect knowledge of any organism’s genetics and ecology, in the past as well as in the present.

Experiments with tractable model organisms evolving in controlled laboratory environments minimize many of these complications and uncertainties^{15,16}. Moreover, new methods have made it feasible to sequence complete genomes from evolution experiments with bacteria^{17–20}. To date, such analyses have focused on finding the mutations responsible for particular adaptations. However, the application of comparative genome sequencing to experimental

evolution studies also offers the opportunity to address major conceptual issues, including whether the dynamics of genomic and adaptive evolution are coupled very tightly or only loosely^{10,12,13,21,22}.

Genome dynamics and adaptation

To examine the tempo and mode of genomic evolution, we sequenced the genomes of *E. coli* clones sampled at generations 2,000, 5,000, 10,000, 15,000, 20,000 and 40,000 from an asexual population that evolved with glucose as a limiting nutrient for almost 20 years as part of a long-term experiment. The complete sequence of the ancestral strain served as a reference for identifying mutations in the evolved clones, which we refer to by their generation abbreviations 2K, 5K, 10K, 15K, 20K and 40K.

Figure 1 shows all mutations identified in the evolved clones through 20,000 generations. The 45 mutations in the 20K clone include 29 single-nucleotide polymorphisms (SNPs) and 16 deletions, insertions and other polymorphisms (DIPs). Figure 2 shows that the number of mutational differences between the ancestral and evolved genomes accumulated in a near-linear fashion over this period. Any deviation from linearity was not statistically significant based on randomization tests.

The near-linearity of the trajectory for genomic evolution is rather surprising, given that such constancy is widely taken as a signature of neutral evolution¹², whereas the fitness trajectory for this population²³ shows profound adaptation that is strongly nonlinear. In particular, the rate of fitness improvement decelerates over time (Fig. 2), which indicates that the rate of appearance of new beneficial mutations is declining, their average benefit is becoming smaller, or both. These effects, in turn, should cause the rate of genomic evolution to decelerate.

To understand this point, consider a simple model of the substitution of beneficial mutations in a clonal population of haploid organisms. A beneficial mutation has an initial frequency of 1/*N*,

¹Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan 48824, USA. ²Industrial Biotechnology and Bioenergy Research Center, Korea Research Institute of Bioscience and Biotechnology, Yuseong, Daejeon 305-806, Korea. ³Department of Computer Science and Engineering, Chungnam National University, Yuseong, Daejeon 305-764, Korea. ⁴21C Frontier Microbial Genomics and Applications Center, Yuseong, Daejeon 305-806, Korea. ⁵Institut Jean Roget, Laboratoire Adaptation et Pathogénie des Microorganismes, CNRS UMR 5163, Université Joseph Fourier, Grenoble 1, BP 170, F-38042 Grenoble cedex 9, France. ⁶Functional Genomics Program, School of Science, University of Science and Technology, Yuseong, Daejeon 305-333, Korea.

*These authors contributed equally to this work.

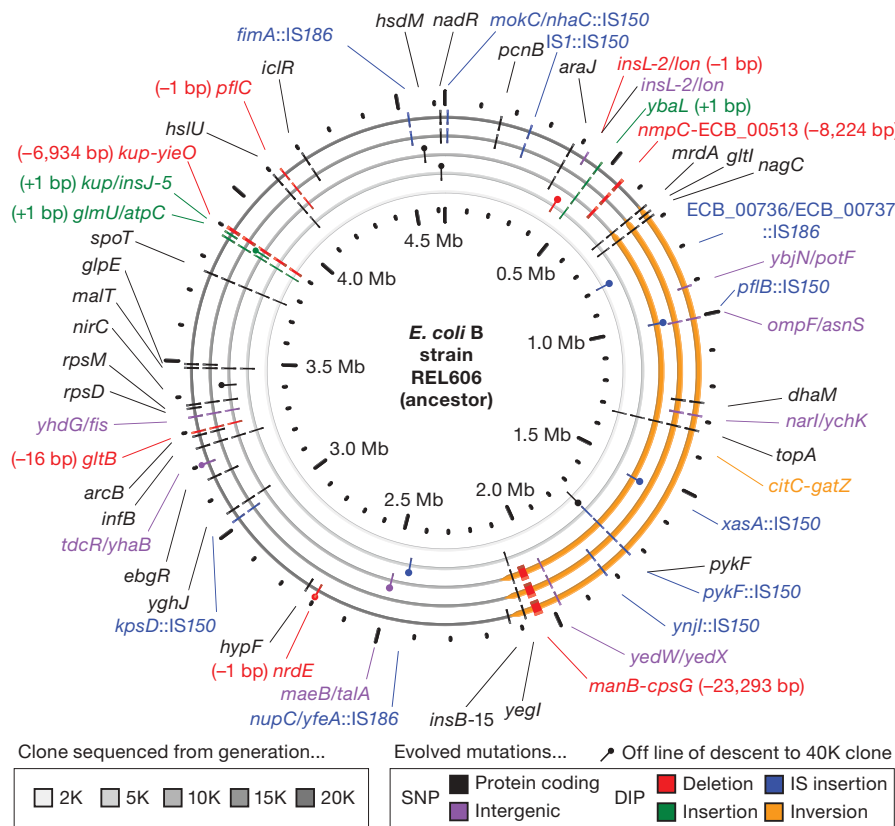


Figure 1 | Mutations found by sequencing genomes sampled between 2,000 and 20,000 generations from an evolution experiment with *E. coli*. The outermost ring represents the genome sampled at 20,000 generations, and labels all genes with SNP mutations in coding (black) and intergenic (purple) regions, and those with DIP mutations including deletions (red), insertions (green), insertion sequence (IS) element insertions (blue), and an inversion between *citC* and *gatZ* (orange). Insertion sequences are transposable elements present in bacterial genomes. The next four rings,

from outer to inner, show mutations present in genomes sampled at 15,000, 10,000, 5,000, and 2,000 generations. The innermost circle shows the genome position and scale in megabase pairs (Mb). Mutations that are off the line of descent to a genome sampled at 40,000 generations are capped with a circle. Only one mutation (*kup/insJ-5*), a 1-base-pair (bp) insertion near an IS150 element, shows an aberrant homoplasic distribution, being present in clones 10K and 20K but not 15K. Precise molecular details for all mutations are shown in Supplementary Tables 1 and 2.

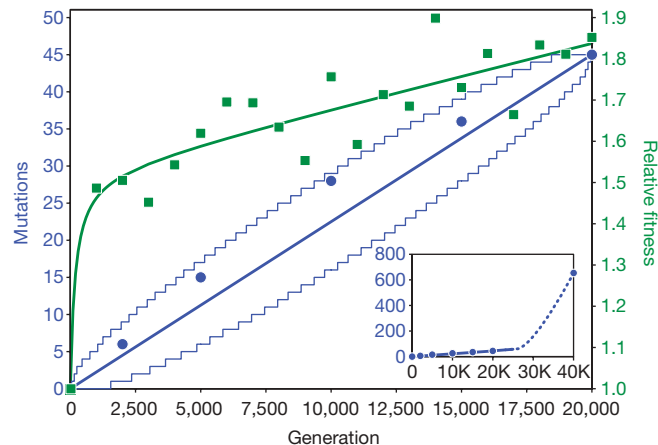


Figure 2 | Rates of genomic evolution and fitness improvement. Blue circles show the total number of genomic changes relative to the ancestor in each sampled clone. The blue line represents a model where mutations accumulate uniformly over time. The light blue curves define the 95% confidence interval for this linear model. Green squares show the improvement of this population's mean fitness relative to the ancestor over time, and the green curve is a hyperbolic plus linear fit of this trajectory. Each fitness estimate is the mean of three assays; most of the spread of points around the fitness trajectory reflects statistical uncertainty inherent to the assays. The inset shows the number of mutations in the 40,000-generation clone; the dashed curve approximates the change in the time course of genomic evolution after a mutator phenotype appeared by about generation 26,500.

where N is population size, and it confers a selective advantage S over its progenitor. Nevertheless, there is some probability that the mutant is lost by drift while it is rare. Given large N , small S and a Poisson distribution of offspring²⁴, a beneficial mutant has a probability of escaping extinction of $\sim 2S$. If the mutant survives, it takes $\tau \approx \log_2(0.5N)/S$ cell generations to increase to 50% frequency. These dynamics thus have two phases. In the first, a population waits for the appearance of a beneficial mutation that avoids extinction by drift with an expected waiting time of $\omega \approx 1/(2SNv)$, where v is the beneficial mutation rate. In the second phase, the mutant spreads by selection, becoming the new majority type after τ generations.

We can explore the relationship between rates of adaptation and genomic evolution under three scenarios. In the first, the substitution of any beneficial mutation has no effect on either the selection coefficient, S , or the beneficial mutation rate, v . The rates of fitness improvement and genome evolution should therefore be constant over the long term. Under the second scenario, the number of possible sites for beneficial mutations is finite, so that v declines with increasing prior substitutions. The expected wait for a beneficial mutation becomes progressively longer, and the trajectories for adaptation and genomic evolution should decelerate in parallel. In the third scenario, the advantage of new beneficial mutations declines as fitness increases. The waiting and sweep times are both inversely proportional to S , so the total expected time between substitutions is also inversely proportional to S . The rate of fitness gain will decelerate with the reduced rate of beneficial substitutions as well as their declining effects, although the trajectories may not be parallel. Under all three scenarios, this model thus predicts declining rates

of both adaptive and genomic evolution or, alternatively, no deceleration in either trajectory.

Predominance of beneficial substitutions

The simplest hypothesis that could explain the discrepancy between the nearly constant rate of genomic change and the sharply decelerating fitness trajectory posits that only a small fraction of all substitutions are beneficial, whereas most are neutral or nearly so^{12,14}. Accordingly, the beneficial substitutions would be concentrated in the early phase of rapid adaptation to the conditions of the experiment, but over time that initial burst would be swamped by the constant accumulation of neutral mutations by drift. However, four lines of evidence allow us to reject this explanation.

First, under this drift hypothesis, one expects disproportionately more synonymous than non-synonymous mutations, because the former have no effect on protein sequence and thus are more likely to be neutral. In fact, all 26 point mutations we found in coding regions (22 in clone 20K, and 4 off the line of descent) are non-synonymous. The probability of observing no synonymous substitutions is only 0.07% if the same base changes were distributed randomly in the coding regions of the ancestral genome.

Second, if mutations had spread by random drift, we would not expect to see mutations in the same genes in the other independently evolved populations of the long-term experiment, because only ~1% of the >4,000 genes in *E. coli* harbour mutations in the population studied here. By contrast, selection should target the same genes in the replicate lines because they started from the same ancestor and evolved in identical environments. Fourteen genes in which mutations were found in our study population have been sequenced in all the other populations after 20,000 generations. There is substantial parallelism, with three cases where all eleven other populations have substituted mutations in the same gene, nine additional genes with mutations in other lines, and only two cases where no other line has a mutation in the same gene (Table 1). In almost all cases, the evolved alleles differ between the populations, so accidental cross-contamination cannot explain these parallel changes.

Third, under the drift hypothesis, we would expect many mutations in individual clones that did not become fixed in the population as a whole. However, almost all mutations in the earlier clones were present in clones from all subsequent generations. For example, four of the six mutations in clone 2K are present in all later clones, and all thirty-four mutations in clone 15K occur in clones 20K and 40K. Moreover, two of the thirteen mutations through 20K that are off the line of descent to the 40K clone occur in genes (*pykF* and *nadR*) where different mutations arose and were substituted later. Both of these genes also have substitutions in all of the other populations, so even these early unsuccessful alleles were probably beneficial, but were nonetheless eliminated because competing sub-lineages had even more beneficial mutations^{25,26}.

Fourth, strains with these mutations should have no fitness advantage under the neutral drift hypothesis. To date, isogenic

strains with ancestral and derived alleles have been constructed at nine loci. In all but one case, the derived allele confers a significant advantage in competition (Table 2). The exception (*ompF*) might also be beneficial in combination with other mutations present in the genetic background in which it evolved, especially because parallel mutations arose in other populations (Table 1). By contrast, another study found that none of 26 random insertion mutations conferred a significant advantage in the same environment²⁷.

Other explanations for rate discordance

Taken together, these four lines of evidence demonstrate that discordance in rates of genomic and adaptive evolution in this experiment cannot be explained by assuming a preponderance of neutral substitutions. Another plausible explanation for the disparity is an ecological one. Fitness levels were measured, at all generations, in competition with the ancestor. In an evolution experiment with yeast, non-transitive ecological interactions gave rise to complex dynamics, such that the cumulative adaptation measured across successive episodes of selection was greater than that measured directly from start to finish²⁸. However, there is no significant discrepancy between the fitness gains summed over shorter intervals and the overall improvement measured from start to finish for the population in our study²³, allowing us to reject this hypothesis.

Clonal interference occurs in asexual organisms when sub-lineages with beneficial mutations are driven extinct by competition with other sub-lineages bearing mutations that are even more beneficial^{25,26}, and this process might contribute to the relatively constant rate of genomic change. In particular, the most beneficial mutations should dominate the early phase of evolution for large populations in a new environment²⁶, but there are more potential mutations that confer small advantages than large ones^{2,13,29,30}. Thus, the supply of contending beneficial mutations may increase enough to sustain a uniform rate of overall genomic change. It may also be relevant that some early substitutions, which contributed the most to fitness improvement, involve global regulatory functions including the stringent response and DNA supercoiling^{31,32}. These mutations have pleiotropic effects on the expression of many genes, and although these changes are beneficial on balance, some of their side effects are probably deleterious. These maladaptive side effects may introduce new opportunities for compensatory changes that restore appropriate expression of other genes and thereby further increase the supply of mutations conferring small advantages.

Emergence of a hypermutable phenotype

Several of the long-term populations evolved mutator phenotypes by 20,000 generations, but the population in our study retained the low ancestral mutation rate to at least that time point^{33,34}. In later generations, however, this population exhibited a greatly elevated rate of genomic evolution (Fig. 2 inset). The 40K genome contains 627 SNP and 26 DIP mutations (Supplementary Tables 3 and 4). As a consequence of the DIP mutations (including six new insertions of

Table 1 | Frequency of parallel mutations in 11 other independently evolved lines

Gene or region	Function	Parallel mutations (%)	Source
<i>nadR</i>	Transcriptional regulator	100	Ref. 42
<i>pykF</i>	Pyruvate kinase	100	Ref. 42
<i>rbs</i> operon	Ribose catabolism	100	Ref. 43
<i>malT</i>	Transcriptional regulator	64	Ref. 44
<i>spoT</i>	Stringent response regulator	64	Ref. 31
<i>mrdA</i>	Cell-wall biosynthesis	45	Ref. 42
<i>infB</i>	Translation initiation factor 2	45*	This study
<i>fis</i>	Nucleoid-associated protein	27	E. Crozat, D.S., unpublished
<i>topA</i>	DNA topoisomerase I	27	E. Crozat, D.S., unpublished
<i>pcnB</i>	Poly(A) polymerase	27	This study
<i>ompF</i>	Outer-membrane porin	18*	This study
<i>rpsD</i>	30S ribosomal protein	18*	This study
<i>rpsM</i>	30S ribosomal protein	0	This study
<i>glmU</i> promoter	Cell-wall biosynthesis	0	M. Stanek, R.E.L., unpublished

* In addition to populations with substitutions, one or more others were polymorphic.

Table 2 | Tests of fitness effect in competition between isogenic constructs

Gene or region	Fitness effect (%)	Significance	Source
<i>topA</i>	13.3	***	Ref. 32
<i>pykF</i> *	11.1	***	D.S., R.E.L., unpublished
<i>spoT</i>	9.4	***	Ref. 31
<i>nadR</i> †	8.1	***	D.S., R.E.L., unpublished
<i>glmU</i> promoter	4.9	***	M. Stanek, T. Cooper, R.E.L., unpublished
<i>fis</i>	2.9	***	Ref. 32
<i>rbs</i> operon†	2.1	***	Ref. 43
<i>malT</i>	0.4	**	Ref. 44
<i>ompF</i> ‡	−9.7	**	D.S., R.E.L., unpublished

* For this mutation, isogenic constructs were made by replacing the evolved allele with the ancestral allele in the evolved genetic background. For all other mutations, isogenic constructs were made in the ancestral background.

† In these two cases, artificial deletions of the genes were constructed in the ancestral background and the fitness effects of those deletions are reported.

‡ The deleterious effect of this mutation could indicate that it hitchhiked to high frequency. Alternatively, its fitness effect was tested only in the ancestral background, and it might be beneficial in association with one or more other evolved alleles.

** $P < 0.01$; *** $P < 0.001$. All significance levels are based on multiple independent competition assays.

IS150, three of IS186 and one of IS1) the genome size of the 40K clone is 4.57×10^6 bp, representing a reduction of 1.2% from the ancestor. Of particular interest is a 1-bp insertion causing a frameshift mutation in the *mutT* gene. Defects in *mutT* specifically cause A·T→C·G transversions³⁵, and 92.3% (553 of 599) of the new point mutations at 40K have this signature, significantly higher than the 23.5% (8 of 34) frequency among earlier mutations (one-tailed Fisher's exact test, $P = 3 \times 10^{-20}$).

We sequenced the site of the *mutT* frameshift in clones from other generations to determine the time-course of its fixation in the population. The mutation occurs in 0 of 3 clones tested at generations 20,000, 25,000, 25,500 and 26,000; 1 of 3 clones at generation 26,500; 2 of 3 clones at generations 27,000, 27,500, 28,000 and 28,500; and 3 of 3 clones at generations 29,000, 29,500, 30,000, 35,000 and 40,000. Thus, the *mutT* mutant appeared by generation 26,500 and soon dominated the population. Luria–Delbrück fluctuation tests^{33,34} indicate that the mutation rate to nalidixic acid resistance increased about 50- to 100-fold in later generations.

The large number of new SNP mutations at 40K is presumably the result of drift coupled with the elevated mutation rate. Unlike the SNP mutations occurring before 20,000 generations, only a small fraction of these new mutations are likely to be beneficial. To test this prediction, we examine below the proportion of synonymous mutations after the mutator phenotype evolved to determine if it is consistent with a random distribution across sites.

Synonymous changes and mutation rates

The fact that no synonymous mutations fixed in the first 20,000 generations is consistent with the low point-mutation rate in *E. coli* and population-genetic theory if those mutations are selectively neutral. According to theory, the expected rate of neutral substitutions equals the rate of neutral mutations¹². We calculate an upper bound for the mutation rate from the Poisson distribution, which specifies a 5% chance that no synonymous substitutions would occur even if three were expected. That upper bound corresponds to a mutation rate of 1.6×10^{-10} per bp per generation given 20,000 generations, a genome length of 4.63×10^6 bp, and the fact that 20.4% of all possible point mutations are synonymous. This inferred rate lies between earlier estimates from mutation studies³⁶ and comparative analyses of *E. coli* and *Salmonella enterica*³⁷.

In the 40K genome, by contrast, 13.9% (83 of 599) of the new base substitutions (those not in the 20K genome) are synonymous. This fraction is lower than would be expected if 20.4% of random substitutions were synonymous. However, mutations in the 40K genome are highly skewed towards A·T→C·G transversions, which have a lower probability of causing synonymous changes than other point mutations. To reflect this mutational bias, we grouped point mutations into two categories: *mutT*, either A→C or T→G; and non-*mutT*, all other base substitutions. These categories have probabilities of synonymous mutations of 11.3% and 22.1%, respectively, and they are represented by 553 and 46 new mutations, respectively, in the 40K

genome. The observation of 83 synonymous substitutions is slightly higher than the random expectation of 71.6 based on the sum of these two binomial distributions, although the excess of synonymous changes is small and only marginally significant at best (one-tailed $P = 0.105$). The small excess of synonymous substitutions implies that a high proportion of late-arising non-synonymous changes are also neutral or nearly so under the conditions of the evolution experiment.

We can also use the number of synonymous substitutions to estimate the point mutation rate after the mutator phenotype evolved. The precise time of origin for the *mutT* subpopulation is unknown, but it was present in the 26,500-generation sample; we assume it arose at generation 25,000 for this estimation. As previously noted, neutral mutations should accumulate at a rate equal to their underlying mutation rate. The lineage leading to the 40K clone accumulated all or almost all of its 83 synonymous substitutions after it became a mutator. Given roughly 15,000 generations, a final genome length of 4.57×10^6 bp, and the fact that only 11.3% of *mutT* point mutations should produce synonymous changes, the 83 such instances imply a point-mutation rate of 1.1×10^{-8} per bp per generation. This rate is about 70-fold higher than the upper bound estimated before the mutator phenotype evolved.

Perspective and outlook

Genome re-sequencing in the context of experimental evolution provides new opportunities for quantifying evolutionary dynamics. We observed discordance between the rates of genomic change and fitness improvement during a 20-year experiment with *E. coli* in two respects. First, mutations accumulated at a near-constant rate even as fitness gains decelerated over the first 20,000 generations. Second, the rate of genomic evolution accelerated markedly when a mutator lineage became established later. The fluid and complex coupling observed between the rates of genomic evolution and adaptation even in this simple system cautions against categorical interpretations about rates of genomic evolution in nature without specific knowledge of molecular and population-genetic processes. Our results also call attention to new opportunities for population-genetic models to explore the long-term dynamic coupling between genome evolution and adaptation, including the effects of clonal interference, compensatory adaptation, and changing mutation rates.

METHODS SUMMARY

Evolution experiment. Twelve *E. coli* populations were propagated at 37 °C for 6,000 days in minimal medium supplemented with limiting glucose at 25 mg l^{−1} by transferring 0.1 ml of culture into 9.9 ml of fresh medium each day^{38,39}. The population designated Ara-1 is the focus of this study. Samples were stored periodically at −80 °C and later revived for sequencing and phenotypic analyses. **Genome re-sequencing.** On the basis of the sequence⁴⁰ of the ancestral strain REL606 (GenBank accession number NC_012967.1), NimbleGen microarray-based comparative genome sequencing⁴¹ was first used to screen for mutations in the 2K and 20K clones. All six evolved clones were then sequenced to >50× coverage using the Illumina 1G platform. Mutations were identified using BRESEQ, a custom computational pipeline. Targeted sequencing was used to

confirm almost all mutations in the 20K and earlier clones, and to find parallel mutations in the other long-term populations.

Mutation trajectory. We performed two randomization tests for deviations from linearity in the rate of genome evolution through 20,000 generations, one based on the cumulative distribution of mutations and the other on a time-weighted test statistic. Each test was performed using three data sets: the total number of mutations, only mutations on the line of descent, and only SNP mutations. None of the one-tailed tests was significant at $P < 0.05$.

Fitness trajectory. Fitness levels were previously measured in competition assays between the Ara-1 population samples and a genetically marked variant of the ancestor²³. We fit these data to a hyperbolic plus linear model: $w = at/(b + t) + ct$, where w is mean fitness, t is time, and a , b and c are free parameters. The inclusion of the linear term c significantly improves the fit relative to a hyperbolic-only (a , b) model ($F_{1,17} = 7.715$, $P = 0.0129$).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 July; accepted 28 August 2009.

Published online 18 October; corrected 29 October 2009 (see full-text HTML version for details).

- Darwin, C. *On the Origin of Species by Means of Natural Selection* (Murray, 1859).
- Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1930).
- Wright, S. in *Proc. 6th Int. Cong. Genet.* 1, 356–366 (1932).
- Gould, S. J. & Eldredge, N. Punctuated equilibrium: the tempo and mode of evolution reconsidered. *Paleobiol.* 3, 115–151 (1977).
- Eldredge, N. *et al.* The dynamics of evolutionary stasis. *Paleobiol.* 31, 133–145 (2005).
- Simpson, G. G. *The Major Features of Evolution* (Columbia Univ. Press, 1953).
- Charlesworth, B., Lande, R. & Slatkin, M. A neo-Darwinian commentary on macroevolution. *Evolution* 36, 474–498 (1982).
- Reznick, D. N., Shaw, F. H., Rodd, F. H. & Shaw, R. G. Evaluation of the rate of evolution in natural populations of guppies (*Poecilia reticulata*). *Science* 275, 1934–1937 (1997).
- Schluter, D. *The Ecology of Adaptive Radiations* (Oxford Univ. Press, 2000).
- Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* 314, 119–121 (2006).
- Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* 105, 7899–7906 (2008).
- Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
- Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Univ. Press, 1991).
- Ohta, T. The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286 (1992).
- Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Rev. Genet.* 4, 457–469 (2003).
- Buckling, A., Maclean, C. R., Brockhurst, M. A. & Colegrave, N. The *Beagle* in a bottle. *Nature* 457, 824–829 (2009).
- Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732 (2005).
- Fiegna, F., Yu, Y. T., Kadam, S. V. & Velicer, G. J. Evolution of an obligate social cheater to a superior cooperator. *Nature* 441, 310–314 (2006).
- Herring, C. D. *et al.* Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nature Genet.* 38, 1406–1412 (2006).
- Hegreness, M. & Kishony, R. Analysis of genetic systems using experimental evolution and whole-genome sequencing. *Genome Biol.* 8, 201 (2007).
- King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116 (1975).
- Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* 392, 917–920 (1998).
- de Visser, J. A. G. M. & Lenski, R. E. Long-term experimental evolution in *Escherichia coli*. XI. Rejection of non-transitive interactions as cause of declining rate of adaptation. *BMC Evol. Biol.* 2, 19 (2002).
- Haldane, J. B. S. A mathematical theory of natural and artificial selection. V. Selection and mutation. *Proc. Camb. Philos. Soc.* 23, 838–844 (1927).
- Muller, H. J. Some genetic aspects of sex. *Am. Nat.* 66, 118–138 (1932).
- Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. *Genetica* 102/103, 127–144 (1998).
- Remold, S. K. & Lenski, R. E. Contribution of individual random mutations to genotype-by-environment interactions in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* 98, 11388–11393 (2001).
- Paquin, C. E. & Adams, J. Relative fitness can decrease in evolving asexual populations of *S. cerevisiae*. *Nature* 306, 368–371 (1983).
- Orr, H. A. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* 56, 1317–1330 (2002).
- Perfeito, L., Fernandes, L., Mota, C. & Gordo, I. Adaptive mutations in bacteria: high rate and small effects. *Science* 317, 813–815 (2007).
- Cooper, T. F., Rozen, D. E. & Lenski, R. E. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* 100, 1072–1077 (2003).
- Crozat, E., Philippe, N., Lenski, R. E., Geiselmann, J. & Schneider, D. Long-term experimental evolution in *Escherichia coli*. XII. DNA topology as a key target of selection. *Genetics* 169, 523–532 (2005).
- Sniegowski, P. D., Gerrish, P. J. & Lenski, R. E. Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387, 703–705 (1997).
- Cooper, V. S. & Lenski, R. E. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* 407, 736–739 (2000).
- Friedberg, E. C., Walker, G. C. & Siede, W. *DNA Repair and Mutagenesis* (ASM Press, 1995).
- Drake, J. W. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl Acad. Sci. USA* 88, 7160–7164 (1991).
- Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA* 96, 12638–12643 (1999).
- Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *Am. Nat.* 138, 1315–1341 (1991).
- Lenski, R. E. Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. *Plant Breed. Rev.* 24, 225–265 (2004).
- Jeong, H. *et al.* Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J. Mol. Biol.* doi:10.1016/j.jmb.2009.09.052 (26 September 2009).
- Albert, T. J. *et al.* Mutation discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nature Methods* 2, 951–953 (2005).
- Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl Acad. Sci. USA* 103, 9107–9112 (2006).
- Cooper, V. S., Schneider, D., Blot, M. & Lenski, R. E. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J. Bacteriol.* 183, 2834–2841 (2001).
- Pelosio, L. *et al.* Parallel changes in global protein profiles during long-term experimental evolution in *Escherichia coli*. *Genetics* 173, 1851–1869 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank collaborators in the Lenski and Schneider laboratories for sharing unpublished data; N. Hajela and J. I. Lee for isolation of genomic DNA; C. T. Brown, C. Epstein, C. H. Lee and J. Plotkin for discussion; N. Hajela, L. Ekinwe and S. Simpson for years of technical assistance with the long-term lines; and W. J. Dittmar for assistance with fluctuation tests. We acknowledge support from the DARPA 'Fun Bio' Program (to R.E.L.); the US National Science Foundation (to J.E.B. and R.E.L.); the Agence Nationale de la Recherche Programme 'Génomique Microbienne à Grande Echelle', Centre National de la Recherche Scientifique, and Université Joseph Fourier (to D.S.); and the 21C Frontier Microbial Genomics and Applications Center Program, Ministry of Education, Science and Technology, Republic of Korea (to J.F.K.).

Author Contributions R.E.L., D.S. and J.F.K. conceived the project and its components. D.S.Y., J.E.B., S.H.Y., H.J., T.K.O. and J.F.K. performed the genome sequencing and confirmatory analyses. D.S. sequenced specific genes in other populations and performed additional molecular procedures. J.E.B. developed code for data analyses and statistical simulations. R.E.L. directs the long-term experiment while J.F.K. directed the genomics work. R.E.L., J.E.B. and J.F.K. analysed the data and wrote the paper. J.E.B., D.S.Y., S.H.Y., R.E.L., D.S. and J.F.K. prepared figures and tables.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.E.L. (lenski@msu.edu) or J.F.K. (jfk@kribb.re.kr).

METHODS

Genomic DNA isolation. Clones 2K (strain REL1164A), 5K (REL2179A), 10K (REL4536A), 15K (REL7177A), 20K (REL8593A) and 40K (REL10938) from the Ara-1 population of the *E. coli* long-term evolution experiment^{38,39} were revived from stocks kept at -80°C in 15% glycerol by growth overnight in LB medium. DNA was harvested and purified from several millilitres of each culture with the Qiagen Genomic-tip 100/G kit.

NimbleGen comparative genome sequencing (CGS). Genomic DNA from the 2K and 20K clones was sent to NimbleGen Systems for comparative genome sequencing. Tiling arrays with 29-mer probes optimized for length, melting temperature, and mismatch position based on the REL606 ancestral genome sequence⁴⁰ were used to detect mutational differences between the ancestor and each evolved clone. Data were visualized using SignalMap version 1.8. All putative SNP sites and candidates for DIP variations (deletions, insertions and inversions) were checked by capillary Sanger sequencing of PCR products amplified directly from the genome. The NimbleGen CGS approach did not find several mutations that had been previously discovered in evolved clones (see later), indicating that a high false-negative rate was an impediment to this approach.

Illumina whole-genome shotgun (WGS) re-sequencing. Genomic DNA samples from all six Ara-1 clones were sequenced on a 1G Genome Analyzer (Illumina) by Macrogen. Standard procedures produced data sets of opposite-strand, mated 36-bp read pairs with inserts averaging ~ 135 bp and calibrated base quality scores. Mutations in each genome were then identified using BRESEQ, a pipeline implemented in Perl for analysing bacterial WGS re-sequencing data (J. E. Barrick, unpublished algorithm). The average coverage at positions with only unique read alignments in each of the six sequenced genomes was between $55\times$ and $65\times$. More than 99.9% of the positions outside of repeated regions in each genome had at least tenfold coverage. These levels are sufficient to ensure with great confidence that SNP mutations at virtually all positions in the reference genome would be discovered if present. For additional verification of mutations predicted by BRESEQ, we also used the software MAQ⁴⁵ (version 0.7.1), which predicted the same set of base substitutions in each WGS re-sequencing data set.

BRESEQ pipeline for mutation discovery. BRESEQ analyses gapped matches between each read sequence and the reference genome produced by MUMmer⁴⁶ (version 3.20) with minimal exact match and extension requirements of 14 bp. For each read, it first determines the set of best alignments that are not contained within other matches and that have fewer mismatches than other alignments with the same endpoints. Of these matches, those that do not contain at least three of the four possible DNA bases or where 80% or more of the length of the match is a single base are eliminated. These nearly homopolymeric matches are typically spurious read sequences. For remaining matches that have a unique best match in the reference genome, internal ambiguities in gapped alignments are systematically re-aligned to consistent reference genome positions by using a Needleman–Wunsch algorithm⁴⁷. The ends of each alignment are also trimmed whenever it is possible that mutations introducing small indels would be compatible with an alternative, equally valid alignment to the reference sequence.

On the basis of unique alignments, BRESEQ predicts base changes and short indels from the calibrated base quality scores that support and contradict each candidate mutation. New sequence junctions (such as those produced by deletions) are predicted from hybrid reads consisting of two regions with best matches to discontinuous sites in the reference genome. The junction prediction procedure involves first assembling a list of candidate junction sequences compatible with these matches, then re-querying hybrid reads against these candidates with MUMmer, and finally predicting consensus junctions from reads that match the new candidates better than they match any portion of the original genome. Additionally, deletions too large to be identified as gaps in individual read alignments are predicted by taking all reference positions with zero unique coverage and propagating outward until the unique coverage at a position exceeds an arbitrary cutoff.

BRESEQ generates HTML output files with the genomic contexts of putative mutations annotated using BioPerl⁴⁸. These tables are linked to alignments of read sequences so that information supporting and contradicting each predicted mutation can be examined in more detail by the user. Additionally, coverage histograms are generated for the entire genome and for each large deletion with the R statistics package⁴⁹. These output files were used to determine the precise extent of deletions and the locations of new sequence junctions. BRESEQ performs a reference-based comparison, rather than *de novo* assembly and therefore this procedure may not reliably detect point mutations, indels and genomic rearrangements involving repeated sequences that occur at several locations in the genome, or short tandem repeats that approach or exceed the 36-bp size of the input reads.

Mutation identification. Supplementary Table 1 shows the precise genomic locations and other details for all 34 SNP mutations found in the genomes of the Ara-1 2K, 5K, 10K, 15K, and 20K clones. Supplementary Table 2 provides details for the 21 DIP mutations identified in these genomes, which include a large inversion, three 1-bp insertions, three 1-bp deletions, four more extensive deletions, and ten IS element insertions.

The following seven SNP mutations were discovered before genome re-sequencing: *mrda* (ref. 42), *ompF/asnS* (D. Schneider, unpublished), *topA* (ref. 32), *yhdG/fis* (ref. 32), *malT* (ref. 44), *spoT* (ref. 31) and 20K *nadR* (ref. 42). Four of these known SNP mutations (*mrda*, *yhdG/fis*, *spoT* and 20K *nadR*) were found independently using NimbleGen CGS, but the other three were not. Of the remaining 23 SNP mutations present in either the 2K or 20K genomes, NimbleGen CGS identified all except five (*insL-2/lon*, *nagC*, 2K *pykF*, *insB-15* and *hypF*). By contrast, analysis of Illumina WGS data identified all seven of the known mutations and 33 of the 34 total SNP mutations reported. The final SNP mutation (*insB-15*) is within a multicopy IS element and was found in this study during additional Sanger sequencing. We also verified that the *kup/insJ-5* 1-bp insertion on the border of an IS150 element shows an aberrant homoplastic distribution, being present in clones 10K and 20K but not 15K, with additional targeted sequencing.

Five DIP mutations were discovered before genome re-sequencing: Δ (*nmpC*-ECB_00513) (ref. 50), *inv(citC-gatZ)* (ref. 50), *pykF::IS150* (ref. 50), *glmU/atpC* (M. Stanek and R.E.L., unpublished), and Δ (*kup-yieO*) (ref. 43). Four of these mutations were found independently using NimbleGen CGS, with the exception being the 1-bp insertion in the *glmU/atpC* intergenic region. Five of the twelve other DIP mutations present in the 2K and 20K genomes were first discovered using NimbleGen CGS (*ynjI::IS150*, Δ (*manB-cpsG*), *kpsD::IS150*, Δ *gltB*, *fimA::IS186*), and the 1-bp insertion in the *kup/insJ-5* intergenic region was found during additional Sanger sequencing. Analysis of Illumina WGS data identified 20 of the 21 DIP mutations reported, all except *inv(citC-gatZ)*. New IS-insertions predicted by WGS re-sequencing were confirmed as size changes in PCR-amplified fragments.

We identified a total of 627 SNP and 26 DIP mutations in the 40K clone genome (Supplementary Tables 3 and 4). In addition to the 15 DIP mutations in the 20K clone that are on the line of descent, the 11 other DIP mutations include four IS-insertions, two 1-bp insertions, one 6-bp insertion, one 1-bp deletion, one 61-bp deletion, and two larger deletions of roughly 7 and 22 kb. Only two mutations that are present in the 20K clone (*tdcR/yhaB* and *nrdE*) are not found at 40K. One of the additional IS-insertions in the 40K clone is an IS186 element in the *nupC/yfeA* intergenic region at the exact site where one occurs in the 5K clone. Because this mutation is missing in all of the sequenced clones from intervening generations, this insertion is the second example of a homoplastic change that evidently originated independently in two sub-lineages. Also, one of the new 1-bp insertions adds another G directly adjacent to the previous *kup/insJ-5* 1-bp insertion. We did not further confirm most of the predicted mutations in the 40K genome owing to the large number, but their quality is on par with the mutations that were correctly predicted in the earlier clones from WGS data.

Parallel mutations. We PCR-amplified and sequenced the *infB*, *pcnB*, *ompF*, *rpsD* and *rpsM* genes of three clones isolated at 20,000 generations from each of the 11 other experimental populations. Mutations were counted as parallel changes if they were within the protein coding sequence or upstream promoter elements, but not if they were downstream of the reading frame.

Fluctuation tests. We performed Luria–Delbrück fluctuation tests³³ to confirm that the Ara-1 population evolved an elevated mutation rate. Bacteria were revived from frozen stocks by growth overnight in LB medium. After dilution and 24 h of re-growth in Davis minimal medium supplemented with 25 mg l^{-1} glucose, we inoculated 24 replicate 10-ml cultures of Davis minimal medium with 250 mg l^{-1} glucose with 100–1,000 cells. After 24 h of growth to stationary phase, these cultures were concentrated by centrifugation and plated on LB agar containing $20\text{ }\mu\text{g ml}^{-1}$ nalidixic acid. The mutation rates to resistance of the mixed populations archived at 20,000, 30,000 and 40,000 generations were estimated as 5.8×10^{-10} , 1.7×10^{-8} and 6.3×10^{-9} per cell division by the maximum likelihood method using a custom Perl script. These values are roughly 4, 120 and 45 times the mutation rate of 1.4×10^{-10} per cell division estimated in the same experimental block for the ancestral strain.

Tests of constant rate of genomic evolution. We examined three data sets consisting of the total number of SNP and DIP mutations in the 2K, 5K, 10K, 15K and 20K genomes (TOT: 6, 15, 28, 36, 45), only those changes on the line of descent to the 40K clone (LOD: 4, 12, 22, 36, 43), and only the SNP mutations in each genome (SNP: 3, 9, 16, 22, 29). The LOD subset represents the estimated rate at which mutations fixed in the population; it excludes the effects of sampling within-population variation, which are unknown and may fluctuate over time given the effect of selective sweeps in purging variation. The SNP subset

represents those mutations that are most commonly modelled in theoretical studies of evolutionary biology. We tested the hypothesis that the rate of mutation accumulation was linear in two ways.

Cumulative distribution test. We used a randomization test to determine whether there was evidence that the number of mutations observed in the 2K, 5K, 10K, 15K and 20K genomes was inconsistent with an underlying probability distribution giving a linear increase over time. For each data set, 10 million simulated data sets were generated by randomly redistributing 20K mutations with a uniform probability at times between 0 and 20,000 generations. The confidence intervals shown in Fig. 1 and also *P*-values for deviations from linearity at a given generation were determined directly from these simulated distributions of mutation number over time. This test is similar in design to the Kolmogorov–Smirnov test, but it takes into account the uneven spacing of the observations of mutation number with respect to time and enforces stricter confidence limits on observations near the ends of the cumulative distribution. If any point from the observed genomic data set fell outside of the 95% confidence interval then the null hypothesis of linearity would be rejected. The most extreme deviations are not significant even by a one-tailed test for observing more mutations than would be expected from the uniform distribution for any of the three data sets: TOT (10K, *P* = 0.052), LOD (15K, *P* = 0.092), SNP (5K, *P* = 0.227).

Time-weighted test statistic. We tested the apparent skew towards excess mutations in early generations using a time-weighted test statistic equal to the summation of the number of new mutations observed in each subsequent genome multiplied by the generation at which that genome was sampled. For example, the value of this test statistic for the TOT data set is 487,000 (equal to $6 \times 2,000 + 9 \times 5,000 + 13 \times 10,000 + 8 \times 15,000 + 9 \times 20,000$). We determined the significance of the observed value of the test statistic for each data set using a randomization test implemented with the STATISTICS101 resampling program (<http://www.statistics101.net/>). The mutations at 20K were redistributed randomly with a uniform probability distribution with respect to time to generate randomized data sets consisting of counts of mutations occurring by 2,000, 5,000, 10,000, 15,000 and 20,000 generations. Then, one-tailed *P*-values were assigned to each data set by comparing the observed value of the test statistic to the distribution of simulated test-statistic values obtained from performing the randomization procedure 10 million times. The skew of excess mutations towards early

generations, as measured by this test statistic, is not significant for any of the three data sets: TOT (*P* = 0.066), LOD (*P* = 0.256), or SNP (*P* = 0.299).

Lack of early synonymous substitutions. In the 2K, 5K, 10K, 15K and 20K clone genomes, 26 different point mutations within protein reading frames were observed, and all of these altered the encoded amino acid. Using the total codon frequencies in the ancestral genome sequence tabulated using a custom Perl script, we calculated the chances that each of the 12 possible nucleotide substitutions would result in a non-synonymous substitution if it occurred at a random coding position (A→C: 0.893; A→G: 0.758; A→T: 0.898; C→A: 0.762; C→G: 0.808; C→T: 0.559; G→A: 0.698; G→C: 0.785; G→T: 0.785; T→A: 0.797; T→C: 0.600; T→G: 0.840). From these probabilities, it is straightforward to calculate the 0.07% probability of observing zero synonymous mutations, by chance alone, given the observed distribution of the 26 coding base substitutions (A→C: 2; A→G: 2; A→T: 3; C→A: 2; C→T: 3; G→A: 7; G→C: 1; G→T: 1; T→A: 1; T→G: 4). The observed excess of non-synonymous substitutions is highly significant.

Synonymous substitution rates. Reading frames annotated in the ancestral genome sequence were extracted and analysed to calculate the probability that a random mutation would be synonymous using a custom Perl script. These calculations assume that the relative rates of different base changes are equal, except as indicated otherwise for the *mutT* mutator.

45. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
46. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
47. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
48. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**, 1611–1618 (2002).
49. R Development Core Team. R: A language and environment for statistical computing Pages (<http://www.R-project.org/>). (R Foundation for Statistical Computing, 2007).
50. Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics* **156**, 477–488 (2000).

ARTICLES

The role of DNA shape in protein–DNA recognition

Remo Rohs^{1*}, Sean M. West^{1*}, Alona Sosinsky^{1†}, Peng Liu¹, Richard S. Mann² & Barry Honig¹

The recognition of specific DNA sequences by proteins is thought to depend on two types of mechanism: one that involves the formation of hydrogen bonds with specific bases, primarily in the major groove, and one involving sequence-dependent deformations of the DNA helix. By comprehensively analysing the three-dimensional structures of protein–DNA complexes, here we show that the binding of arginine residues to narrow minor grooves is a widely used mode for protein–DNA recognition. This readout mechanism exploits the phenomenon that narrow minor grooves strongly enhance the negative electrostatic potential of the DNA. The nucleosome core particle offers a prominent example of this effect. Minor-groove narrowing is often associated with the presence of A-tracts, AT-rich sequences that exclude the flexible TpA step. These findings indicate that the ability to detect local variations in DNA shape and electrostatic potential is a general mechanism that enables proteins to use information in the minor groove, which otherwise offers few opportunities for the formation of base-specific hydrogen bonds, to achieve DNA-binding specificity.

The ability of proteins to recognize specific DNA sequences is a hallmark of biological regulatory processes. The determination of the three-dimensional structures of numerous protein–DNA complexes has provided a detailed picture of binding, revealing a structurally diverse set of protein families that exploit a wide repertoire of interactions to recognize the double-helix¹. Nucleotide sequence-specific interactions often involve the formation of hydrogen bonds between amino-acid side chains and hydrogen-bond donors and acceptors of individual base pairs. It has long been recognized that every base pair has a unique hydrogen-bonding signature in the major groove, but that this is not the case in the minor groove². Thus, the expectation has been that the recognition of specific DNA sequences would take place primarily in the major groove by the formation of a series of amino-acid- and base-specific hydrogen bonds¹. This ‘direct readout’ mechanism is consistent with observations derived from three-dimensional structures of protein–DNA complexes, but it is far from the entire story.

In many complexes, the DNA assumes conformations that deviate from the structure of an ideal B-form double helix^{3–5}, sometimes bending in such a way to optimize the protein–DNA interface⁶, and in some cases undergoing large conformational changes as in the opening of the minor groove in the complex formed between TBP and the TATA box^{7,8}. The term ‘indirect readout’ was coined⁹ to describe such recognition mechanisms that depend on the propensity of a given sequence to assume a conformation that facilitates its binding to a particular protein. The bases involved in such mechanisms need not be in contact with the protein and, for example, can be found in linker sequences that connect two half-sites that are themselves bound by individual protein subunits^{10,11}.

We recently described an example of a new readout mechanism, the recognition of local sequence-dependent minor-groove shape¹² that is distinct from previously described indirect readout mechanisms. In this case, the sequence-dependence of minor-groove width and corresponding variations in electrostatic potential are used by the

Drosophila Hox protein Sex combs reduced (SCR) to distinguish small differences in nucleotide sequence¹². Here we report that this mechanism is a widely used mode of protein–DNA recognition that involves the creation of specific binding sites for positively charged amino acids, primarily arginine, within the minor groove. Minor-groove narrowing is found to be correlated with A-tracts^{13,14}, usually defined as stretches of four or more As or Ts that do not contain the flexible TpA step¹⁵, but extended here to include as few as three base pairs (see later). Our results offer fundamentally new insights into the structural and energetic origins of protein–DNA binding specificity, and thus have important implications for the prediction of transcription-factor-binding sites in genomes.

Arginine is enriched in narrow minor grooves

The percentage of minor-groove contacts associated with each amino acid, classified according to the width of the minor groove, was determined (Fig. 1a). Arginine constitutes 28% of all amino-acid residues that contact the minor groove and is notably enriched in narrow minor grooves, defined here by a groove width of <5.0 Å (compared to 5.8 Å in ideal B-DNA). Remarkably, 60% of the residues in narrow minor grooves are arginines, compared to 22% in minor grooves that are defined as not narrow—that is, width ≥ 5.0 Å. A smaller enrichment is also observed for lysine but the overall population of lysines within the minor groove is much less than for arginine.

Binding to the minor groove is a characteristic of many, but not all, protein superfamilies and a large subset of these contact a narrow minor groove (Table 1). Moreover, if the minor groove is contacted, arginines are likely to be involved, and the likelihood that an arginine will be present becomes even greater for narrow minor grooves (Supplementary Table 1).

We compiled the DNA sequence preferences for protein–DNA complexes in which an arginine contacts a narrow minor groove (Fig. 1b). The figure shows that the base pair that has the shortest contact distance with the arginine guanidinium group has a 78%

¹Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 Saint Nicholas Avenue, New York, New York 10032, USA. ²Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West 168th Street, HHSC 1104, New York, New York 10032, USA. [†]Present address: Institute of Structural and Molecular Biology, School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK.

*These authors contributed equally to this work.

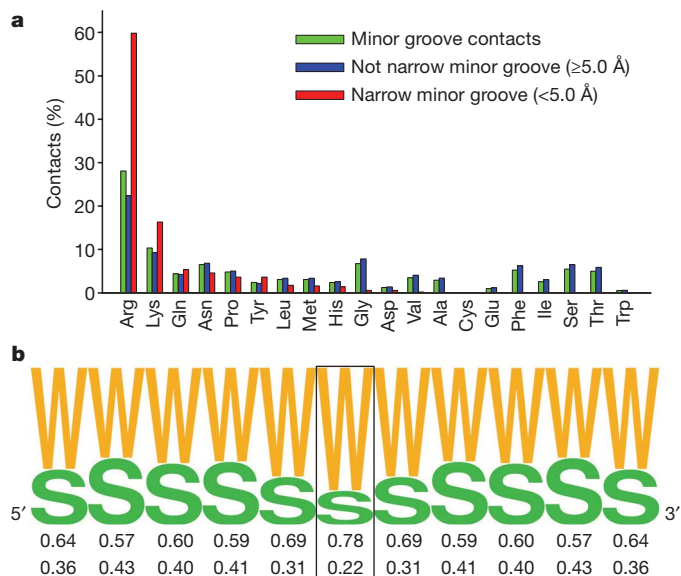


Figure 1 | Amino acid frequencies in minor grooves. **a**, Histograms for each amino acid illustrate the frequency with which they are observed in any minor groove (green), in minor grooves with a width of ≥ 5.0 Å (blue), and in narrow minor grooves of width < 5.0 Å (red). **b**, Frequency of AT (W) and GC (S) base pairs in sequences of 229 sites contacted by arginines in narrow minor grooves. The central base pair (boxed) is contacted by arginine. Frequencies are symmetrized by using both complementary strands.

probability of being an AT and 22% of being a GC. Neighbouring base pairs in both the 5' and 3' directions surrounding the closest contacting base pair also have a strong tendency to be AT. Taken together, these data demonstrate that arginines tend to bind narrow minor grooves in AT-rich DNA.

AT-rich sequences tend to narrow minor grooves

We calculated minor-groove widths for all tetranucleotides contained in Protein Data Bank (PDB) structures for both free DNA (Fig. 2a) and DNA in complexes with proteins (Fig. 2b). There is a large spread of values due in part to end effects and to the effects of crystal packing, but some trends are evident nevertheless. For example, for free DNA structures most of the tetranucleotides with narrow minor grooves (width < 5.0 Å) are AT-rich (Fig. 2a and Supplementary Table 2a).

Table 1 | Protein superfamilies with minor-groove contacts

Narrow minor groove	Not narrow minor groove
SRF-like	DNA repair protein MutS, domain I
IHF-like DNA-binding proteins	Origin of replication-binding domain, RBD-like
Histone-fold	DNA/RNA polymerases
DNA breaking-rejoining enzymes	Eukaryotic DNA topoisomerase I, amino-terminal DNA-binding fragment
Zn2/Cys6 DNA-binding domain	Ribonuclease H-like
Homeodomain-like	TATA-box binding protein-like
p53-like transcription factors	
Lambda repressor-like DNA-binding domains	
Winged helix DNA-binding domain	
Leucine zipper domain	
C-terminal effector domain of the bipartite response regulators	
Restriction endonuclease-like	
Glucocorticoid receptor-like (DNA-binding domain)	

Listed are SCOP superfamilies⁴⁶ that have an arginine minor-groove contact within a distance of < 6.0 Å from the base. Superfamilies that use arginine to contact a narrow minor groove (< 5.0 Å) and those that use arginine to contact a not narrow minor groove (≥ 5.0 Å) are shown. Only superfamilies with a minimum of ten protein chains in PDB structures bound to DNA at least one helical turn long are included. The percentages of chains with minor-groove contacts vary considerably among SCOP superfamilies and are provided in Supplementary Table 1.

Similar behaviour is observed in protein–DNA complexes (Fig. 2b and Supplementary Table 2b). In contrast, tetranucleotides with wide minor grooves have a strong tendency to be GC-rich.

The correlation between AT content and groove width is not unexpected given the fact that A-tracts are known to produce narrow minor grooves. However, TpA steps have a tendency to widen the minor groove¹⁵, so it was of interest to determine whether the distinct properties of A-tracts and TpA steps are reflected in our tetranucleotide data set. We find that 67% of tetranucleotides composed only of AT base pairs have a narrow minor groove, but that this number increases to 82% if we exclude TpA steps so as to consider only A-tracts. Even A-tracts of length three have a strong tendency to narrow the minor groove. Forty-three per cent of the tetranucleotides with a minor groove width of < 5.0 Å have an A-tract length of three, a percentage that decreases to 11% of tetranucleotides with canonical minor-groove widths (between 5.0 and 7.0 Å) and to 4% of tetranucleotides with minor grooves wider than 7.0 Å (Supplementary Fig. 1). Furthermore, compared to other AT-rich sequences, A-tracts are specifically enriched in DNAs with narrow minor grooves (Supplementary Fig. 1). Thus, although A-tracts are usually thought of as requiring four or more base pairs, in part because a minimum of four is required to rigidify the DNA¹⁴, this analysis shows that A-tracts as short as length three are positively correlated with narrow minor grooves.

Arginines recognize enhanced electrostatic potentials

The minor-groove width and electrostatic potential versus binding-site sequence for several complexes whose binding interface includes an arginine inserted into the minor groove is plotted in Fig. 3 and Supplementary Fig. 2. The correlation of width and potential as well as the tendency of arginines to be located close to minima in width and potential is evident. In this section we highlight a few specific examples of how arginine–minor-groove interactions are used in DNA recognition.

Figure 3a represents the ternary complex of the *Drosophila* Hox protein Ultrabithorax (UBX) and its cofactor Extradenticle (EXD) bound to DNA¹⁶. In this complex, Arg 5 of UBX, which is a conserved residue across all homeodomains, inserts into a narrow region formed by a 4-base-pair (bp) A-tract. An example of a long and very narrow A-tract that binds $\alpha 2$ -Arg 7 from the MATa1–MAT $\alpha 2$ complex with DNA is shown (Fig. 3b)¹⁷. In contrast, $\alpha 2$ -Arg 4 inserts into a shallower region at one end of the A-tract, where there are local minima in width and potential that are smaller than at the Arg 7 site in the centre of the A-tract. The two POU domains of the mammalian OCT1 (also known as POU2F1)–PORE complex bind to two A-tracts (Fig. 3c) in which the minima are positioned in such a way as to provide binding sites for four arginines, two from each POU domain¹⁸.

The location of these A-tracts with respect to other nucleotide sequence features can be used to generate specificity, as previously discussed for the Hox protein SCR¹². In the case of SCR binding, the position of a TpA step within an AT-rich region has a critical role in binding specificity. A similar strategy is used by the motility gene repressor (MogR) in which two long A-tracts separated by a TpA step produce two arginine-binding sites¹⁹ (Fig. 3d). The unique shape recognized by these two arginines probably contributes to the position of the MogR-binding site along the DNA sequence. The overall tendency of TpA steps to widen the minor groove is most apparent when they are positioned between two A-tracts (as in SCR¹² and MogR¹⁹) where the TpA step acts as a 'hinge' between more rigid elements^{15,20}. In other contexts, owing to their flexibility, TpA steps can also be accommodated in narrow minor grooves²¹. An example is provided by the bipartite DNA-binding domain of Tc3 transposase in which the arginines bind to a narrow region containing a TATA box²² that displays enhanced negative electrostatic potential (Fig. 3e).

Although less frequent, arginines also bind narrow grooves associated with non-A-tract sequences. Figure 3f summarizes features of the binding of the 434 repressor to its operator²³ that contains 7 bp

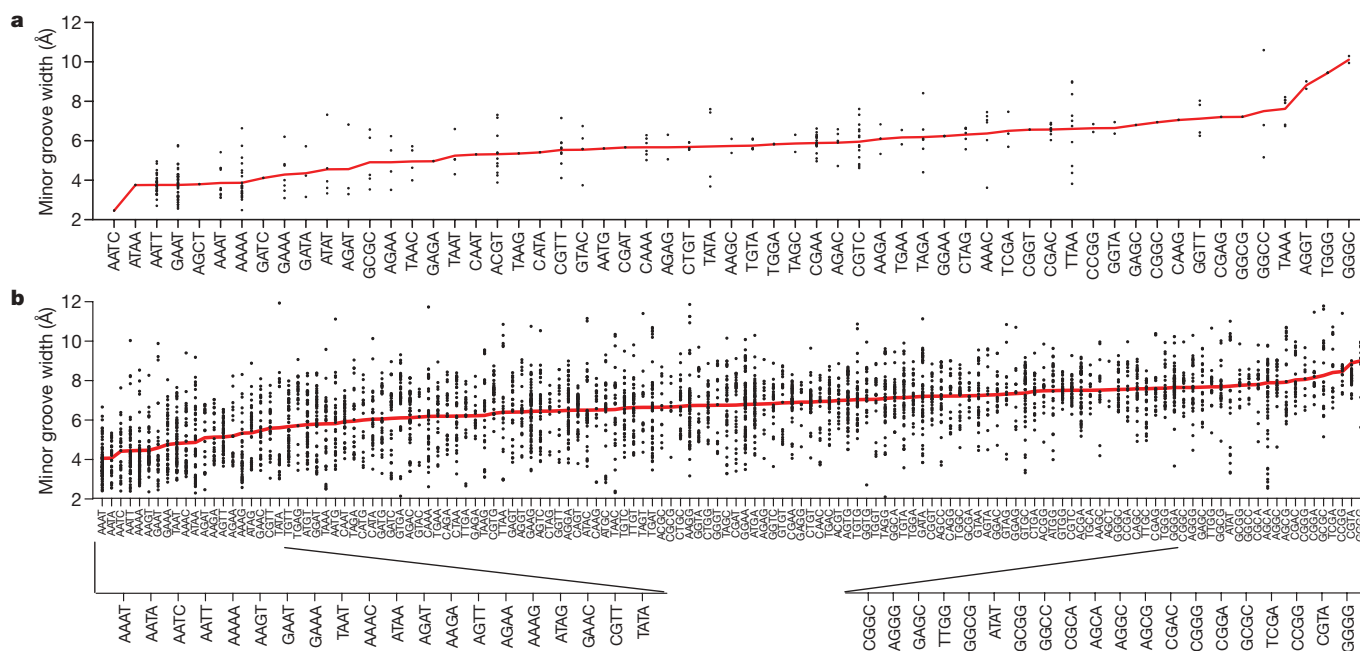


Figure 2 | Distribution of tetranucleotide sequences according to average minor-groove width. Tetranucleotides from structures with a minimum length of one helical turn for which minor-groove width can be defined are ordered by average minor-groove width (red). The widths of all

tetranucleotides are shown (black) and the sequence, average width, and occurrence in our data set are given in Supplementary Table 2. **a**, The 59 unique tetranucleotides from free DNA structures. **b**, The set of all 136 unique tetranucleotides derived from protein–DNA complexes.

that are all AT with the exception of a central CG. (The guanine amino group tends to widen narrow grooves, but a single GC base pair can be accommodated with only little disruption).

Arginine minor-groove interactions in the nucleosome

Figure 4a plots minor-groove width and electrostatic potential along the DNA sequence of the nucleosome core particle containing recombinant histones and a 147-bp DNA fragment (PDB code 1kx5)²⁴. There are 14 minima in minor-groove width corresponding to regions where the DNA bends so as to wrap around the histone core. As earlier, there is a marked correlation between width and potential. The variation in width between the narrowest and widest regions is about 5 Å, and the difference between the maxima and minima in electrostatic potential is about 6 kT e⁻¹ (Fig. 4a). As a consequence, there should be a strong driving force for basic amino acids to bind to narrow regions and indeed arginines are found in 9 of the 14 minima. These arginines are shown in Fig. 4b where the nucleosomal DNA has been colour-coded by minor-groove width. (Although all 14 narrow minor-groove regions are contacted by arginines²⁴ only 9 satisfy our criteria of <6.0 Å between arginine atoms and base atoms in the groove). A similar repeating pattern of narrow minor grooves that are contacted by arginines is seen in all 35 available nucleosome crystal structures (Supplementary Fig. 3a, b).

Because short A-tracts narrow the minor groove and facilitate the bending of DNA, we would expect to see a periodicity of A-tracts in DNA sequences bound by nucleosomes *in vivo*. Previous analyses have focused on dinucleotide statistics^{25,26}, although it has been known for some time that there is a periodic pattern of AAA and AAT trinucleotides in nucleosome core DNA^{27,28}. An analysis of DNA sequences bound *in vivo* by yeast nucleosomes²⁹ reveals a clear periodicity for A-tracts of at least length three (denoted 3+, Fig. 4c). Moreover, nucleosomal DNAs contain, on average, 10.0 A-tracts of length 3+ (Fig. 4d). Periodicity is also detected for A-tracts of length 4+ and even 5+, although the number per nucleosome decreases to 4.1 and 1.6, respectively (Supplementary Fig. 3). Thus, even though long A-tracts tend to be excluded from the nucleosome³⁰, A-tracts of ≤5 bp, when present, are used to facilitate bending of the DNA around the histone core.

To evaluate the effect of TpA steps, we compared the periodicities of A-tracts of length three to those of other trinucleotides composed only of AT base pairs. Trinucleotides that contain TpA steps have a much weaker periodic signal than A-tracts of length three, which exclude the TpA step (Supplementary Fig. 4). Together, this analysis suggests that many of the sequence periodicities in nucleosomal DNA reflect the presence of short A-tracts that lead to narrow regions in the minor groove, which are in turn recognized by a complementary set of arginines present on the surface of the nucleosome core particle.

Effects of groove width on electrostatic potential

The remarkable correlation between minor-groove width and electrostatic potential (Figs 3 and 4) is primarily due to the properties of the Poisson–Boltzmann equation that have been extensively discussed in the literature³¹. Biological macromolecules are less polarizable than the aqueous solvent and, in the language of classical physics, can be thought of as a low dielectric region embedded in a high dielectric solvent. Solutions of the Poisson–Boltzmann equation for DNA showed that contours of electrostatic potential owing to backbone phosphates follow the shape of the DNA and that the potentials are the most negative within the grooves³². This effect is due to electrostatic focusing, first observed for the protein superoxide dismutase³¹, where the narrow active site focuses electric field lines away from the protein and into the high dielectric solvent. The same physical phenomenon produces enhanced potentials in grooves, accounting for the strong correlation described earlier.

To establish the source of the effect in quantitative terms, we calculated the potentials for the MogR-binding site¹⁹ when the dielectric constant is set to 80 both inside the macromolecule and in the solvent (Fig. 5, dashed line) and for the case where the two dielectric constants are different (Fig. 5, solid line). Notably, a large enhancement of electrostatic potentials is only observed when the dielectric constant of the macromolecule and solvent are different, reflecting the focusing of electric field lines described qualitatively earlier. The small effect seen when the dielectric constant is the same results from the phosphates being closer to the centre of the groove when it is narrow (see Supplementary Fig. 5 for a breakdown of the contributions to the net electrostatic potential). Both sets of calculations were

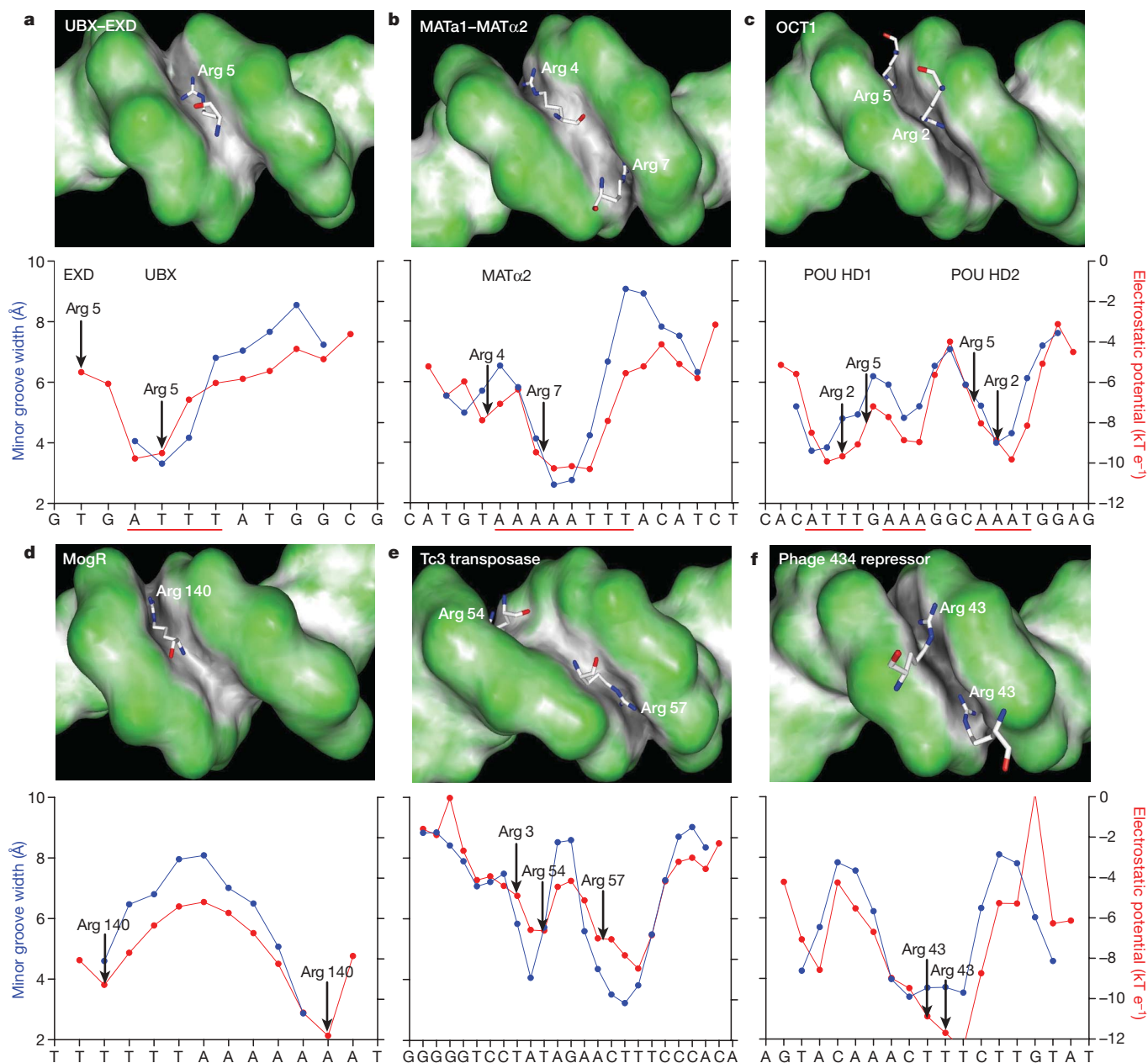


Figure 3 | Specific examples of minor-groove shape recognition by arginines. a–f, DNA shapes of the binding sites of UBX–EXD (PDB code 1b8i)¹⁶ (a), MATa1–MAT α 2 (PDB code 1akh)¹⁷ (b), and OCT1–PORE (PDB code 1hf0)¹⁸ (c), the MogR repressor (PDB code 3fdq)¹⁹ (d), the Tc3 transposase (PDB code 1u78)²² (e) and the phage 434 repressor (PDB code 2or1)²³ (f) are shown in GRASP surface representations^{31,47}, with convex

surfaces colour-coded in green and concave surfaces in grey/black. Plots of minor-groove width (blue) and electrostatic potential in the centre of the minor groove (red) are shown below. Arginine contacts (defined by the closest distance between the guanidinium groups and the bases) are indicated. A-tract sequences are highlighted by a solid red line, the TATA box in e by a dashed line.

carried out at physiological salt concentrations. Although ionic strength influences the absolute values of the potentials, the dielectric boundary effect remains essentially the same (Supplementary Fig. 6).

Why are arginines preferred over lysines?

It is surprising that there is a substantial population of arginines in the minor groove and a large enrichment when the groove is narrow, whereas the effects for lysines are more modest (Fig. 1a). Arginines have been known for some time to be enriched relative to lysines in protein–protein³³ and protein–DNA³⁴ interfaces, and the difference has generally been attributed to the ability of the guanidinium group to engage in more hydrogen bonds than the amino group of lysine³⁵. To evaluate this idea we determined the number of hydrogen bonds formed by all the arginines and lysines in our data set that penetrate the minor groove. Surprisingly, on average, less than one hydrogen

bond is formed by either amino-acid side chain to DNA (0.9 for arginine and 0.6 for lysine), and the standard deviations are such that this difference is insignificant (Supplementary Table 3).

An alternative explanation derives from the difference in the size of the cationic moieties of the two residues. According to the classical Born model, the solvation free energies of ions are proportional to the inverse of their radii³¹, suggesting that it is energetically less costly to remove a charged guanidinium group from water than it is to remove the smaller amino group of a lysine. To test this quantitatively, we calculated the change in free energy in transferring arginine and lysine from water to a medium of dielectric constant 2 (see Methods for details). The difference in the transfer free energies between the two residues ranges from 2.3 to 6.6 kcal mol^{−1}, depending on the force field that was used, with lysine consistently having the higher value (Supplementary Table 4). These results indicate that

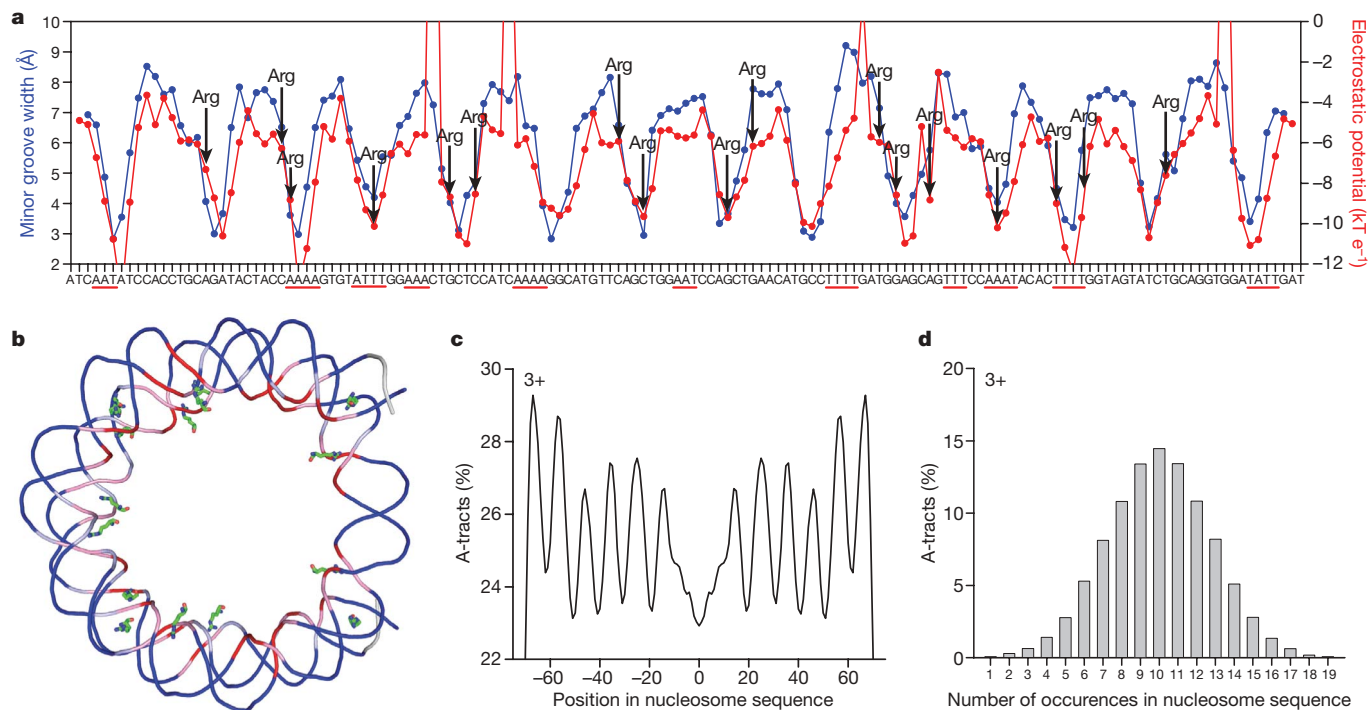


Figure 4 | Minor-groove shape recognition in the nucleosome.

a, Correlation of minor-groove width of the nucleosome core particle (PDB code 1kx5)²⁴ (blue) and electrostatic potential (red). Arginine contacts (defined by the closest distance between the guanidinium groups and the bases) are indicated. A-tract sequences are highlighted by solid red lines. **b**, Schematic representation of the DNA backbone in the nucleosome

the higher prevalence of arginines compared to lysines in minor grooves is due, at least in part, to the greater energetic cost of removing a charged lysine from water than removing a charged arginine.

Concluding remarks

We have shown that there is a marked enrichment of arginines in narrow regions of the DNA minor groove that provides the basis for a new DNA recognition mechanism that is used by many families of DNA-binding proteins. A readout mechanism on the basis of groove width requires a connection between sequence and shape. This connection seems to be provided in part by A-tracts, which have a strong tendency to narrow the groove, producing binding sites for arginines that, when spaced appropriately on the protein surface, offer a complementary set of positive charges that can recognize local variations in shape. Arginines often insert into the minor groove as part of short sequence motifs (for example, Arg-Gln-Arg in the Hox protein SCR¹², Arg-Lys-Lys-Arg in POU homeodomains¹⁸, Arg-Pro-Arg in

colour-coded by minor-groove width (red ≤ 4.0 Å, pink >4.0 Å and ≤ 5.0 Å, light blue >5.0 Å and ≤ 6.0 Å, dark blue >6.0 Å), including all arginines that contact the minor groove. **c**, The distribution of A-tracts of length 3 bp or longer in 23,076 yeast nucleosome-bound DNA sequences²⁹. **d**, Histogram of the occurrence of A-tracts of length three or longer in the same data set²⁹.

Engrailed³⁶, Arg-Gly-His-Arg in MATA1–MAT α 2 (ref. 17), Arg-Arg-Gly-Arg in the nuclear orphan receptor³⁷ and Arg-Gly-Gly-Arg in the human orphan receptor³⁸), thus offering a variety of presentation modes that can contribute to the specificity of DNA shape recognition.

The tendency of A-tracts to narrow the minor groove is primarily due to their ability to assume conformations, by propeller twisting, that lead to the formation of inter-base-pair hydrogen bonds in the major groove¹⁵. This network is disrupted by TpA steps as notably seen in the MogR-binding site¹⁹. GC base pairs also have a tendency to widen the minor groove. The combination of these and other factors, such as the effects induced by flanking bases that are not directly located within the binding site³⁹, can produce a complex minor-groove landscape that offers numerous possibilities for specific interactions with proteins. Indeed, minor-groove geometry is no doubt the result of the interaction of intrinsic and protein-induced structural effects.

The physical mechanisms described here are markedly evident in the nucleosome. The energetic cost of narrowing and bending the DNA in regions where the backbone faces inward will be reduced by the presence of short A-tracts that have an intrinsic propensity to assume such conformations and hence to bend the DNA²⁸. Furthermore, the penetration of arginines into the minor groove at sites where the DNA bends and the groove is narrow^{21,40} provides an important stabilizing interaction.

The variations in DNA shape observed in protein–DNA complexes often reflect conformational preferences of free DNA^{4,10,41}. Sequence-dependent conformational preferences have also been observed in computational studies^{11,21,42} and, most recently, analysis of hydroxyl radical cleavage patterns shows that DNA shape is under evolutionary selection⁴³. Such observations indicate that the role of DNA shape must be taken into consideration when annotating entire genomes and predicting transcription-factor-binding sites. The biophysical insights described here, together with the increased availability of high-throughput binding data, offer the hope of major progress in understanding how proteins recognize specific DNA sequences and in the development of improved predictive algorithms.

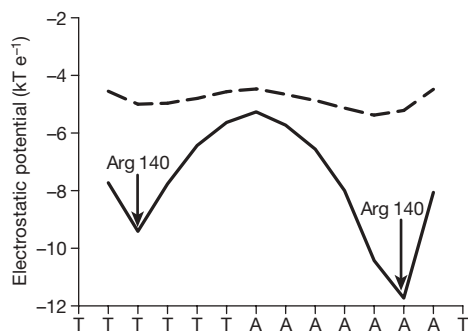


Figure 5 | The biophysical origins of the negative potential of narrow minor grooves. Electrostatic potential in the minor groove of the MogR-binding site (PDB code 3fdq)¹⁹, calculated in the presence of a dielectric boundary ($\epsilon = 2$ in solute and $\epsilon = 80$ in solvent, solid line) and in the absence of a boundary ($\epsilon = 80$ in both solute and solvent, dashed line).

METHODS SUMMARY

Minor-groove geometry was analysed with Curves⁴⁴ for all 1,031 crystal structures of protein–DNA complexes in the PDB that have any amino acid contacting base atoms. Protein side chains contact the minor groove in 69% of those structures that have at least one helical turn of DNA. The probabilities for each amino acid to contact the minor groove were calculated for three groups of DNAs: total, narrow and not narrow. Proteins were grouped on the basis of 40% sequence identity. The properties of free DNAs and DNAs bound to proteins were analysed on the basis of the minor-groove widths of tetranucleotides, defined at the central base-pair step.

All 35 crystal structures of the nucleosome available in the PDB were analysed. The analysis of nucleosomal DNA is based on 23,076 sequences in an *in vivo* yeast data set²⁹. The signal for a sequence motif in nucleosomal DNA is positive for a base pair when the base pair comprises any part of the sequence motif. Frequencies were symmetrized by analysing both complementary DNA strands.

Electrostatic potentials were obtained from solutions to the non-linear Poisson–Boltzman equation at physiological ionic strength using the DelPhi program^{31,45}. Regions inside the molecular surface of the DNA were assigned a dielectric constant of 2, whereas the solvent was assigned a value of 80. The potential is reported at a reference point at the centre of the minor groove. The reference point is located close to the bottom of the groove in approximately the plane of a base pair. This definition provides a measure of electrostatic potential as a function of base sequence. Solvation free energies of amino acids were calculated for extended conformations of arginine and lysine side chains and compared for four different force fields.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 July; accepted 2 September 2009.

- Garvie, C. W. & Wolberger, C. Recognition of specific DNA sequences. *Mol. Cell* **8**, 937–946 (2001).
- Seeman, N. C., Rosenberg, J. M. & Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA* **73**, 804–808 (1976).
- Travers, A. A. DNA conformation and protein binding. *Annu. Rev. Biochem.* **58**, 427–452 (1989).
- Shakked, Z. *et al.* Determinants of repressor/operator recognition from the structure of the *trp* operator binding site. *Nature* **368**, 469–473 (1994).
- Lu, X. J., Shakked, Z. & Olson, W. K. A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300**, 819–840 (2000).
- Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **359**, 505–512 (1992).
- Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512–520 (1993).
- Kim, J. L., Nikolov, D. B. & Burley, S. K. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**, 520–527 (1993).
- Otwinowski, Z. *et al.* Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature* **335**, 321–329 (1988).
- Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D. & Shakked, Z. DNA bending by an adenine–thymine tract and its role in gene regulation. *Proc. Natl Acad. Sci. USA* **98**, 8490–8495 (2001).
- Rohs, R., Sklenar, H. & Shakked, Z. Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* **13**, 1499–1509 (2005).
- Joshi, R. *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**, 530–543 (2007).
- Burkhoff, A. M. & Tullius, T. D. Structural details of an adenine tract that does not cause DNA to bend. *Nature* **331**, 455–457 (1988).
- Haran, T. E. & Mohanty, U. The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* **42**, 41–81 (2009).
- Crothers, D. M. & Shakked, Z. in *Oxford Handbook of Nucleic Acid Structures* (ed. Neidle, S.) 455–470 (Oxford Univ. Press, 1999).
- Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. *Nature* **397**, 714–719 (1999).
- Li, T., Jin, Y., Vershon, A. K. & Wolberger, C. Crystal structure of the MATa1/MATα2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res.* **26**, 5707–5718 (1998).
- Reményi, A. *et al.* Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. *Mol. Cell* **8**, 569–580 (2001).
- Shen, A., Higgins, D. E. & Panne, D. Recognition of AT-Rich DNA binding sites by the MgrR repressor. *Structure* **17**, 769–777 (2009).
- Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V. & Feigon, J. DNA A-tract bending in three dimensions: solving the dA₄T₄ vs. dT₄A₄ conundrum. *Proc. Natl Acad. Sci. USA* **101**, 1177–1182 (2004).
- Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K. & Zhurkin, V. B. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.* **371**, 725–738 (2007).
- Watkins, S., van Pouderooyen, G. & Sixma, T. K. Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res.* **32**, 4306–4312 (2004).
- Aggarwal, A. K., Rodgers, D. W., Drott, M., Ptashne, M. & Harrison, S. C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899–907 (1988).
- Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097–1113 (2002).
- Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA* **77**, 3816–3820 (1980).
- Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
- Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191**, 659–675 (1986).
- Travers, A. A. & Klug, A. in *DNA Topology and its Biological Effects* (eds Cozzarelli, N. R. & Wang, J. C.) 57–106 (Cold Spring Harbor Press, 1990).
- Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput. Biol.* **4**, e1000216 (2008).
- Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19**, 65–71 (2009).
- Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **268**, 1144–1149 (1995).
- Jayaram, B., Sharp, K. A. & Honig, B. The electrostatic potential of B-DNA. *Biopolymers* **28**, 975–993 (1989).
- Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53–64 (1997).
- Nadassy, K., Wodak, S. J. & Janin, J. Structural features of protein–nucleic acid recognition sites. *Biochemistry* **38**, 1999–2017 (1999).
- Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res.* **29**, 2860–2874 (2001).
- Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: a framework for understanding homeodomain–DNA interactions. *Cell* **63**, 579–590 (1990).
- Meinke, G. & Sigler, P. B. DNA-binding mechanism of the monomeric orphan nuclear receptor NGFI-B. *Nature Struct. Biol.* **6**, 471–477 (1999).
- Gearhart, M. D., Holmbeck, S. M., Evans, R. M., Dyson, H. J. & Wright, P. E. Monomeric complex of human orphan estrogen related receptor-2 with DNA: a pseudo-dimer interface mediates extended half-site recognition. *J. Mol. Biol.* **327**, 819–832 (2003).
- Rohs, R., West, S. M., Liu, P. & Honig, B. Nuance in the double-helix and its role in protein–DNA recognition. *Curr. Opin. Struct. Biol.* **19**, 171–177 (2009).
- Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003).
- Locasale, J. W., Napoli, A. A., Chen, S., Berman, H. M. & Lawson, C. L. Signatures of protein–DNA recognition in free DNA binding sites. *J. Mol. Biol.* **386**, 1054–1065 (2009).
- Tolstorukov, M. Y., Virnik, K. M., Adhya, S. & Zhurkin, V. B. A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.* **33**, 3907–3918 (2005).
- Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324**, 389–392 (2009).
- Lavery, R. & Sklenar, H. Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.* **6**, 655–667 (1989).
- Rocchia, W. *et al.* Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.* **23**, 128–137 (2002).
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
- Petrey, D. & Honig, B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* **374**, 492–509 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by National Institutes of Health (NIH) grants GM54510 (R.S.M.) and U54 CA121852 (B.H. and R.S.M.). The authors thank A. Califano for many helpful conversations.

Author Contributions R.R., A.S., R.S.M. and B.H. contributed to the original conception of the project; S.M.W. and R.R. generated and analysed the data assisted by P.L.; and R.R., S.M.W., R.S.M. and B.H. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.H. (bh6@columbia.edu) or R.S.M. (rsm10@columbia.edu).

METHODS

Calculation of minor-groove width. There were in total 1,031 crystal structures of protein–DNA complexes in the PDB as of 1 June 2008, in which the DNA was contacted by any amino-acid side chain at a distance <6.0 Å from base atoms. Of these structures, 567 contained at least one helical turn, and no chemical modifications or deformations that prevent the calculation of minor-groove width. Groove geometry was analysed using Curves⁴⁴ and minor-groove width was calculated as a function of base sequence by averaging all the Curves levels given for each nucleotide.

Statistical analysis of protein–DNA contacts. Of the 567 protein–DNA structures in our data set, 392 have at least one minor-groove contact defined by a distance of <6.0 Å between any base and side-chain atoms. To avoid an over-sampling bias, proteins in this data set that shared $\geq 40\%$ sequence identity were grouped to create 109 groups. The average number of contacts within each group was subsequently averaged over all 109 groups. These averages were divided by the sum of the average number of contacts for all amino acids to calculate the total minor-groove contacts, contacts in not narrow minor grooves (≥ 5.0 Å), and contacts in narrow minor grooves (<5.0 Å), for each amino acid.

Hydrogen-bond contacts between amino-acid side chains and the DNA bases and phosphates, water molecules and other protein atoms were identified with the HBplus program⁴⁸.

Structural annotation of DNA-binding proteins. The proteins in our data set of protein–DNA complexes were classified in SCOP⁴⁶ superfamilies. Proteins for which SCOP annotations were not available were annotated manually or using the ASTRAL database⁴⁹.

Correlation of tetranucleotide structure and sequence. Tetranucleotides in free DNA and protein–DNA complexes were used to analyse the base sequence propensity of minor-groove regions as a function of minor-groove width. The minor-groove width of a tetranucleotide was defined by the average of all Curves⁴⁴ levels for groove width of the second nucleotide and the first level of the third nucleotide, which describes groove width at the central base-pair step. End regions and irregular tetranucleotides were excluded by requiring groove width definitions for at least one Curves level of each of the four nucleotides. Tetranucleotides from nucleosomal DNA were excluded from this analysis because the DNA is strongly deformed and the spacing between narrow regions is fixed at about one helical turn, thus adding a bias to the results. When applied to the 521 protein–DNA complexes in our data set, these criteria allowed the analysis of all 136 possible unique tetranucleotides. When applied to the 88 free

DNA structures in our data set, the same criteria resulted in the analysis of 59 unique tetranucleotides. To increase coverage for the free DNA data set, NMR structures were included if dipolar coupling data were used in the refinement.

Propensity of sequence motifs in nucleosomes. The structural analysis of nucleosomes includes all 35 crystal structures in the PDB as of 1 May 2009. The sequence analysis was based on 23,076 nucleosome sequences of length 146–148 bp in a yeast *in vivo* data set²⁹. These nucleosome sites were scanned for sequence motifs such as A-tracts of different lengths, TpA steps, or other AT-rich regions. A given motif contributed to a positive signal for any base pair that overlapped that motif, thus longer motifs contributed signals to more nucleotide positions. The frequencies of all motifs were symmetrized by analysing both complementary strands.

Calculations of electrostatic potentials. Electrostatic potentials were obtained from solutions to the non-linear Poisson–Boltzman equation at 0.145 M salt using the DelPhi program^{31,45}. Partial charges and atomic radii were taken from the Amber force field⁵⁰. The interior of the molecular surface of the solute molecule (calculated with a 1.4 Å probe sphere) was assigned a dielectric constant of $\epsilon = 2$, whereas the exterior aqueous phase was assigned a value of $\epsilon = 80$. Debye–Hückel boundary conditions and five focusing steps were used with a cubic grid size of 165 (a grid size of 185 was used for the nucleosome).

The electrostatic potential is reported at a reference point close to the bottom of the minor groove approximately in the plane of base pair i . The reference point i is defined as the geometric midpoint between the O4' atoms of nucleotide $i + 1$ in the 5'–3' strand, and nucleotide $i - 1$ in the 3'–5' strand¹². Where the DNA strongly bends into the major groove the reference point can clash with the guanine amino group and cause large positive potentials (as seen in Fig. 4a for three regions of the nucleosome).

Desolvation free energies were calculated with the DelPhi program^{31,45} for the transfer of arginine and lysine side chains in extended conformations from water to a medium of dielectric constant $\epsilon = 2$. Transfer free energies were calculated for each of the two side chains based on charge distributions and atomic radii taken from Amber⁵⁰ and three other force fields (see Supplementary Table 4).

48. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).

49. Brenner, S. E., Koehl, P. & Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28**, 254–256 (2000).

50. Cornell, W. D. *et al.* A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

LETTERS

A γ -ray burst at a redshift of $z \approx 8.2$

N. R. Tanvir¹, D. B. Fox², A. J. Levan³, E. Berger⁴, K. Wiersema¹, J. P. U. Fynbo⁵, A. Cucchiara², T. Krühler^{6,7}, N. Gehrels⁸, J. S. Bloom⁹, J. Greiner⁶, P. A. Evans¹, E. Rol¹⁰, F. Olivares⁶, J. Hjorth⁵, P. Jakobsson¹¹, J. Farihi¹, R. Willingale¹, R. L. C. Starling¹, S. B. Cenko⁹, D. Perley⁹, J. R. Maund⁵, J. Duke¹, R. A. M. J. Wijers¹⁰, A. J. Adamson¹², A. Allan¹³, M. N. Bremer¹⁴, D. N. Burrows², A. J. Castro-Tirado¹⁵, B. Cavanagh¹², A. de Ugarte Postigo¹⁶, M. A. Dopita¹⁷, T. A. Fatkhullin¹⁸, A. S. Fruchter¹⁹, R. J. Foley⁴, J. Gorosabel¹⁵, J. Kennea², T. Kerr¹², S. Klose²⁰, H. A. Krimm^{21,22}, V. N. Komarova¹⁸, S. R. Kulkarni²³, A. S. Moskvitin¹⁸, C. G. Mundell²⁴, T. Naylor¹³, K. Page¹, B. E. Penprase²⁵, M. Perri²⁶, P. Podsiadlowski²⁷, K. Roth²⁸, R. E. Rutledge²⁹, T. Sakamoto²¹, P. Schady³⁰, B. P. Schmidt¹⁷, A. M. Soderberg⁴, J. Sollerman^{5,31}, A. W. Stephens²⁸, G. Stratta²⁶, T. N. Ukwatta^{8,32}, D. Watson⁵, E. Westra⁴, T. Wold¹² & C. Wolf²⁷

Long-duration γ -ray bursts (GRBs) are thought to result from the explosions of certain massive stars¹, and some are bright enough that they should be observable out to redshifts of $z > 20$ using current technology^{2–4}. Hitherto, the highest redshift measured for any object was $z = 6.96$, for a Lyman- α emitting galaxy⁵. Here we report that GRB 090423 lies at a redshift of $z \approx 8.2$, implying that massive stars were being produced and dying as GRBs ~ 630 Myr after the Big Bang. The burst also pinpoints the location of its host galaxy.

GRB 090423 was detected by the Burst Alert Telescope (BAT) on NASA's Swift satellite⁶ at 07:55:19 UT on 23 April 2009. Observations with Swift's X-ray Telescope (XRT), which began 73 s after the trigger, revealed a variable X-ray counterpart and localized its position to a precision of 2.3 arcsec (at the 90% confidence level). Ground-based optical observations in the *r*, *i* and *z* filters starting within a few minutes of the burst revealed no counterpart at these wavelengths (Supplementary Information).

The United Kingdom Infrared Telescope (UKIRT), Hawaii, began imaging about 20 min after the burst, in response to an automated request, and provided the first infrared (2.15- μ m) detection of the GRB afterglow. In parallel, observations in other near-infrared (NIR) filters using the Gemini North 8-m telescope, Hawaii, showed that the counterpart was only visible at wavelengths greater than about 1.2 μ m (Fig. 1). In this range, the afterglow was relatively bright and exhibited a shallow spectral slope, $F_\nu \propto \nu^{-0.26}$, in contrast to the deep limit on any flux at 1.02 μ m. Later observations from Chile using the MPI/ESO 2.2-m telescope, Gemini South and the Very Large Telescope (VLT) confirmed this finding. Such a sharp spectral break cannot be produced by dust absorption at any redshift, and is a

textbook case of a short-wavelength 'drop-out' source. The full grizYJHK spectral energy distribution (SED) obtained ~ 17 h after burst gives a photometric redshift of $z = 8.06^{+0.21}_{-0.28}$, assuming a simple intergalactic medium (IGM) absorption model. Complete details of our imaging campaign are given in Supplementary Table 1.

Our first NIR spectroscopy was performed with the European Southern Observatory (ESO) 8.2-m VLT, starting about 17.5 h after the burst. These observations revealed a flat continuum that abruptly disappeared at wavelengths less than about 1.13 μ m, confirming the origin of the break as being due to Lyman- α absorption by neutral

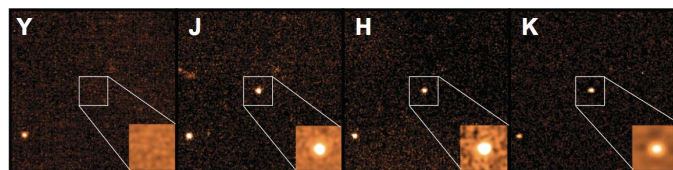


Figure 1 | Multiband images of the afterglow of GRB 090423. The right-most panel shows the discovery image made using the UKIRT Wide Field Infrared Camera with the K filter (centred at 2.15 μ m) at a mid-time of about 30 min after the burst. The other three images (Y, 1.02 μ m; J, 1.26 μ m; H, 1.65 μ m) were obtained approximately 1.5 h after the burst using Gemini North's Near Infrared Imager and Spectrometer (NIRI). The main panels are 40 arcsec to a side, oriented with north to the top and east to the left. Insets, regions around the GRB, smoothed and at higher contrast. The absence of any flux in Y implies a power-law spectral slope between Y and J steeper than $F_\nu \propto \nu^{-1.8}$ and, coupled with the blue colour at longer wavelengths ($J-H(AB) \approx 0.15$ mag), immediately implies a redshift greater than about 7.8 for GRB 090423.

¹Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK. ²Department of Astronomy & Astrophysics, Pennsylvania State University, University Park, Pennsylvania 16802, USA. ³Department of Physics, University of Warwick, Coventry CV4 7AL, UK. ⁴Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. ⁵Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, 2100 Copenhagen, Denmark. ⁶Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstraße 1, 85740 Garching, Germany. ⁷Universe Cluster, Technische Universität München, Boltzmannstrasse 2, 85748 Garching, Germany. ⁸NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ⁹Department of Astronomy, University of California, Berkeley, California 94720-3411, USA.

¹⁰Astronomical Institute "Anton Pannekoek", University of Amsterdam, PO Box 94249, 1090 GE Amsterdam, The Netherlands. ¹¹Centre for Astrophysics and Cosmology, Science Institute, University of Iceland, Dunhagi 5, 107 Reykjavík, Iceland. ¹²Joint Astronomy Centre, 660 North A'ohoku Place, University Park, Hilo, Hawaii 96720, USA. ¹³School of Physics, University of Exeter, Stocker Road, Exeter EX4 4QL, UK. ¹⁴H. H. Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK. ¹⁵Instituto de Astrofísica de Andalucía del Consejo Superior de Investigaciones Científicas, PO Box 03004, 18080 Granada, Spain. ¹⁶European Southern Observatory, Casilla 19001, Santiago 19, Chile. ¹⁷Research School of Astronomy & Astrophysics, The Australian National University, Cotter Road, Weston Creek, Australian Capital Territory 2611, Australia. ¹⁸Special Astrophysical Observatory, Nizhny Arkhyz, Karachai-Circassian Republic, 369167, Russia. ¹⁹Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. ²⁰Thüringer Landessternwarte Tautenburg, Sternwarte 5, 07778 Tautenburg, Germany. ²¹CRESST and NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ²²Universities Space Research Association, 10211 Wincopin Circle, Suite 500, Columbia, Maryland 21044, USA. ²³Department of Astronomy, California Institute of Technology, MC 249-17, Pasadena, California 91125, USA. ²⁴Astrophysics Research Institute, Liverpool John Moores University, Birkenhead CH41 1LD, UK. ²⁵Department of Physics and Astronomy, Pomona College, Claremont, California 91711, USA. ²⁶ASI Science Data Center, Via Galileo Galilei, 00044 Frascati, Italy. ²⁷Department of Physics, Oxford University, Keble Road, Oxford OX1 3RH, UK. ²⁸Gemini Observatory, Hilo, Hawaii 96720, USA. ²⁹Physics Department, McGill University, 3600 Rue University, Montreal, Quebec H3A 2T8, Canada. ³⁰The UCL Mullard Space Science Laboratory, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK. ³¹The Oskar Klein Centre, Department of Astronomy, Stockholm University, 106 91 Stockholm, Sweden. ³²The George Washington University, Washington DC 20052, USA.

hydrogen, with a redshift of $z \approx 8.2$. The spectrum and broadband photometric observations, plotted over model data, are shown in Fig. 2. To obtain a more quantitative estimate of the redshift, we fit the spectra in redshift versus $\log[N_{\text{H I}} (\text{cm}^{-2})]$ space, assuming a flat prior likelihood value for $\log[N_{\text{H I}} (\text{cm}^{-2})]$ of between 19 and 23, which is broadly consistent with the distribution observed for lower-redshift GRB hosts^{7–9}. We take the neutral fraction of the IGM to be 10%, although our conclusions depend only weakly on this assumption. We find the redshift from ISAAC spectroscopy to be $z = 8.19^{+0.03}_{-0.06}$. An additional spectrum, recorded ~ 40 h after the burst using the VLT's Spectrograph for INTEGRAL Field Observations in the Near Infrared confirms this analysis, yielding $z = 8.33^{+0.06}_{-0.11}$ (Supplementary Information). Fitting simultaneously to both spectra and the photometric data points gives our best estimate of the redshift, $z = 8.23^{+0.06}_{-0.07}$. The low signal-to-noise ratio means we that are unable to detect metal absorption features in either spectrum—which would provide a more precise value of the redshift—and prevents a meaningful attempt to measure the IGM H I column density in this instance. Our three independent redshift measures are consistent with that reported from a low-resolution spectrum obtained with the Telescopio Nazionale Galileo, La Palma¹⁰.

The X-ray and NIR light curves of GRB 090423 (Fig. 3) show a broken power-law decay, with evidence of flares in both the X-ray and the infrared bands. The spectral energy distribution is consistent with the presence of the cooling break between the X-ray and optical bands. Apart from the unusually shallow spectral slope of the continuum at wavelengths greater than $1.2 \mu\text{m}$, its afterglow properties in general appear to be consistent with the bulk GRB population (see Supplementary Information for further discussion).

With the standard cosmological parameters (Hubble parameter, $H_0 = 71 \text{ km s}^{-1} \text{ Mpc}^{-1}$; total matter density, $\Omega_{\text{M}} = 0.27$; dark-energy density, $\Omega_{\Lambda} = 0.73$) a redshift of $z = 8.2$ corresponds to a time of only 630 Myr after the Big Bang, when the Universe was just 4.6% of its current age. GRB 090423's inferred isotropic equivalent energy, $E_{\text{iso}} = 1 \times 10^{53} \text{ erg}$ (8–1,000 keV)¹¹, indicates that it was a bright, but not extreme, GRB. Thus, we find no evidence of exceptional behaviour that might indicate an origin in a population III progenitor. First-generation stars are thought more likely to collapse into particularly massive black holes, which in turn may produce unusually long-lived GRBs¹²; this seems not to be the case for GRB 090423.

Indeed, we note that the γ -ray duration of GRB 090423, $t_{90} = 10.3 \text{ s}$, corresponds in the rest frame to only 1.1 s, and the peak energy measured by BAT, 49 keV, is moderately hard in the rest frame. Two other GRBs with $z > 5$ (GRB 060927 and GRB 080913) had similarly short rest-frame durations, leading to some debate¹³ as to whether their progenitors were similar to those of the 'short-hard' class of GRBs, which are not thought to be directly related to core collapse. However, in the case of GRB 090423, a more careful extrapolation of the observed γ -ray and X-ray light curves to lower redshifts shows that its duration would have appeared significantly longer than suggested by naive time-dilation considerations¹⁴. In any event, short GRBs probably have their origins in compact objects that are themselves the end products of massive stars, so the above conclusions will hold irrespective of the population from which GRB 090423 derives.

It has long been recognized that GRBs have the potential to be powerful probes of the early Universe¹⁵. Their association with individual stars means that they serve as a signpost of star formation, even if their host

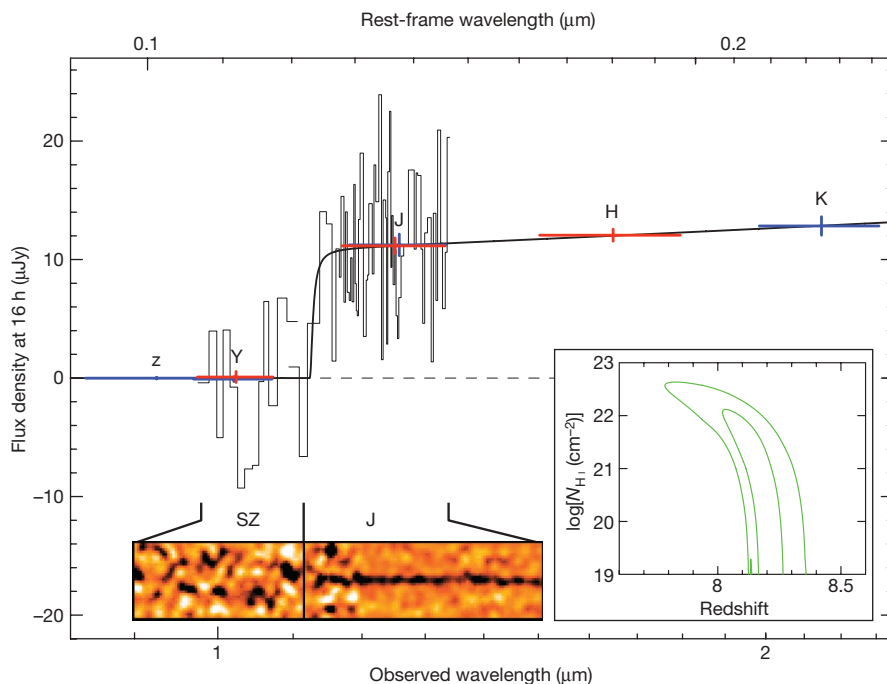


Figure 2 | The composite infrared spectrum of the GRB 090423 afterglow. SZ-band (0.98–1.1 μm) and J-band (1.1–1.4 μm) one- and two-dimensional spectra obtained with the VLT using the Infrared Spectrometer And Array Camera (ISAAC). Also plotted are the sky-subtracted photometric data points obtained using Gemini North's NIRI (red) and the VLT's High Acuity Wide field K-band Imager and Gemini South's Gemini Multi-Object Spectrograph (blue) (scaled to 16 h after the burst and expressed in microjanskys; $1 \text{ Jy} = 10^{-26} \text{ W m}^{-2} \text{ Hz}^{-1}$). The vertical error bars show the 2σ (95%) confidence level, and the horizontal lines indicate the widths of the filters. The shorter-wavelength measurements are non-detections, and emphasize the tight constraints on any transmitted flux below the break. The break itself, at an observed wavelength of about $1.13 \mu\text{m}$, is seen to occur close to the short-wavelength limit of the J-band spectrum, below which,

although noisy, the spectrum shows no evidence of any detected continuum. Details of the data-reduction steps and adaptive binning used to construct these spectra are given in Supplementary Information. A model spectrum showing the H I damping wing for a host galaxy with a hydrogen column density of $N_{\text{H I}} = 10^{21} \text{ cm}^{-2}$ at a redshift of $z = 8.23$ is also plotted (solid black line), and provides a good fit to the data. Inset, allowing for a wider range in possible host $N_{\text{H I}}$ values gives the 1σ (68%) and 2σ confidence contours shown. The fact that no deviation is seen from a power-law spectrum at wavelengths greater than $1.2 \mu\text{m}$, together with its shallow spectral slope, suggests that there is little or no dust along the line of sight through the GRB host galaxy (unless it is 'grey'), consistent with the galaxy being relatively unevolved, and having a low abundance of metals.

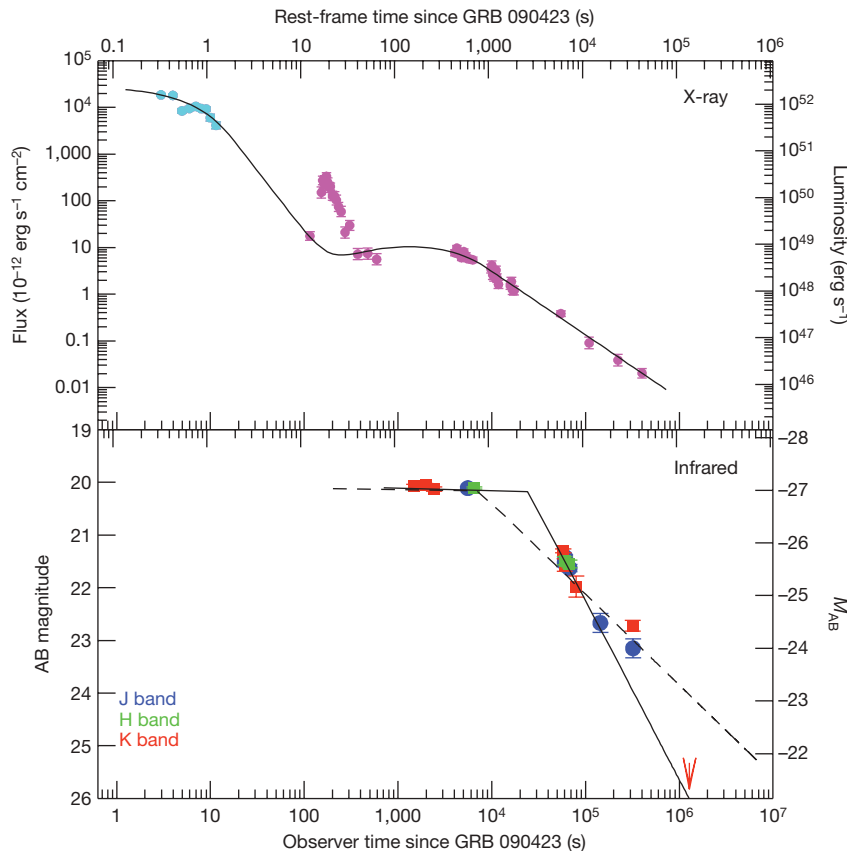


Figure 3 | The X-ray and infrared light curves of GRB 090423. The axes show both observed (left-hand and bottom axes) and rest-frame (right-hand and top axes) quantities. The X-ray light curve was obtained using Swift's BAT (cyan) and XRT (magenta), where the BAT observations have been extrapolated into the X-ray band. The fitted function represents a phenomenological model²⁸ of the prompt and afterglow components. The infrared light curve was obtained using UKIRT, Gemini North, the MPI/ESO 2.2-m telescope and the VLT. For consistency, although individual bands are plotted, they have been transformed into absolute magnitudes in the J band by means of the best-fitting SED ($F_{\nu} \propto \nu^{-0.26}$). We show two illustrative fits to the infrared light curve. The solid line shows a plateau, breaking at

24,000 s to a steeper slope proportional to $\sim t^{-1.4}$. This underestimates the late time points, which must then be interpreted as a flare. The dashed line shows an alternative model, in which mid-time points at $\sim 60,000$ s are instead interpreted as a flare; this is more consistent with the later time points and the X-ray break time at the end of the plateau. However, in this case the post-break slope, proportional to $\sim t^{-0.7}$, is much slower than the X-ray decay at comparable times, and it further requires an additional break in the light curve to accommodate the late-time upper limit. Error bars are 1σ (68% confidence level) and the absolute magnitude scale corresponds to absolute AB magnitudes at $0.136 \mu\text{m}$. See Supplementary Information for further details.

galaxies are too faint to detect directly. Equally important, precise determination of the hydrogen Lyman- α absorption profile can provide a measure of the neutral fraction of the IGM at the location of the burst^{16–20}. With multiple GRBs at redshifts of $z > 7$, and the associated information about the IGM, we could therefore trace the process of reionization from its early stages²¹.

The high redshift of GRB 090423 has several crucial implications. Predictions based on extrapolating the global star-formation-rate density suggest that the observed rate of GRBs at $z \approx 8$ should be about 40% of that at $z \approx 6$ (ref. 12). Given the extra difficulty of identifying afterglows at higher redshifts, our finding is broadly consistent with these predictions. This is extremely encouraging for the prospects of future initiatives aimed at finding high-redshift GRBs and using them to locate and study primordial galaxies and measure the history of star formation at early times^{22–24}. Furthermore, it is close to the redshift range in which the bulk of the cosmic reionization is thought to have taken place^{25–27}. Very high-redshift GRBs for which infrared spectroscopy was possible earlier, or which had brighter afterglows, would provide a direct probe of the progress of reionization. Finding such events is not an unreasonable hope: the most extreme GRBs have had afterglows that were intrinsically significantly brighter than that of GRB 090423 at the same rest-frame time^{3,4}, and our first spectra were recorded more than 15 h after the burst. Spectroscopy with a high signal-to-noise ratio would also provide a measure of the metallicity of the host galaxy, which potentially offers important clues to the

nature of any earlier generations of stars. Because the massive stars that yield GRBs are also likely to belong to the same population that is responsible for reionization, this suggests that GRBs will ultimately be used to constrain both sides—supply and demand—of the cosmic ionization budget in the early Universe.

Received 3 June; accepted 19 August 2009.

1. Woosley, S. E. & Bloom, J. S. The supernova gamma-ray burst connection. *Annu. Rev. Astron. Astrophys.* **44**, 507–556 (2006).
2. Lamb, D. Q. & Reichart, D. E. Gamma-ray bursts as a probe of the very high redshift universe. *Astrophys. J.* **536**, 1–18 (2000).
3. Racusin, J. L. *et al.* Broadband observations of the naked-eye γ -ray burst GRB 080319B. *Nature* **455**, 183–188 (2008).
4. Bloom, J. S. *et al.* Observations of the naked-eye GRB 080319B: implications of nature's brightest explosion. *Astrophys. J.* **691**, 723–737 (2009).
5. Iye, M. *et al.* A galaxy at a redshift $z = 6.96$. *Nature* **443**, 186–188 (2006).
6. Gehrels, N. *et al.* The Swift Gamma-Ray Burst Mission. *Astrophys. J.* **611**, 1005–1020 (2004).
7. Jakobsson, P. *et al.* H I column densities of $z > 2$ Swift gamma-ray bursts. *Astron. Astrophys.* **460**, L13–L17 (2006).
8. Chen, H.-W., Prochaska, J. X. & Gnedin, N. Y. A new constraint on the escape fraction in distant galaxies using γ -ray burst afterglow spectroscopy. *Astrophys. J.* **667**, L125–L128 (2007).
9. Fynbo, J. P. U. *et al.* Low-resolution spectroscopy of gamma-ray burst optical afterglows: biases in the Swift sample and characterization of the absorbers. Preprint at (<http://arxiv.org/abs/0907.3449>) (2009).
10. Salvaterra, R. *et al.* GRB 090423 at a redshift of $z \approx 8.1$. *Nature* doi:10.1038/nature08459 (this issue).

11. von Kienlin, A. *et al.* GRB 090423: Fermi GBM observation (correction of isotropic equivalent energy). *GCN Circ.* **9251** (2009).
12. Bromm, V. & Loeb, A. High-redshift gamma-ray bursts from population III progenitors. *Astrophys. J.* **642**, 382–388 (2006).
13. Zhang, B. *et al.* Physical classification scheme of cosmological gamma-ray bursts and their observational characteristics: on the nature of $z=6.7$ GRB 080913 and some short/hard GRBs. *Astrophys. J.* (in the press); preprint at (<http://arxiv.org/abs/0902.2419v1>) (2009).
14. Zhang, B.-B. & Zhang, B. GRB 090423: pseudo burst at $z=1$ and its relation to GRB 080913. *GCN Circ.* **9279** (2009).
15. Wijers, R. A. M. J. *et al.* Gamma-ray bursts from stellar remnants - probing the universe at high redshift. *Mon. Not. R. Astron. Soc.* **294**, L13–L17 (1998).
16. Miralda-Escude, J. Reionization of the intergalactic medium and the damping wing of the Gunn-Peterson Trough. *Astrophys. J.* **501**, 15–22 (1998).
17. Barkana, R. & Loeb, A. Gamma-ray bursts versus quasars: Ly α signatures of reionization versus cosmological infall. *Astrophys. J.* **601**, 64–77 (2004).
18. Totani, T. *et al.* Implications for cosmic reionization from the optical afterglow spectrum of the gamma-ray burst 050904 at $z = 6.3$. *Publ. Astron. Soc. Jpn* **58**, 485–498 (2009).
19. Greiner, J. *et al.* GRB 080913 at redshift 6.7. *Astrophys. J.* **693**, 1610–1620 (2009).
20. Faucher-Giguere, C.-A., Lidz, A., Hernquist, L. & Zaldarriaga, M. Evolution of the intergalactic opacity: implications for the ionizing background, cosmic star formation, and quasar activity. *Astrophys. J.* **688**, 85–107 (2008).
21. McQuinn, M. *et al.* Probing the neutral fraction of the IGM with GRBs during the epoch of reionization. *Mon. Not. R. Astron. Soc.* **388**, 1101–1110 (2008).
22. Grindlay, J. in *Gamma-Ray Burst: Sixth Huntsville Symposium* (eds Meegan, C., Kouveliotou, C. & Gehrels, N.) 18–24 (AIP Conf. Ser. 1113, American Institute of Physics, 2009).
23. Tanvir, N. R. & Jakobsson, P. Observations of GRBs at high redshift. *Phil. Trans. R. Soc. A* **365**, 1377–1384 (2007).
24. Berger, E. *et al.* Hubble Space Telescope and Spitzer observations of the afterglow and host galaxy of GRB 050904 at $z = 6.295$. *Astrophys. J.* **665**, 102–106 (2007).
25. Komatsu, E. *et al.* Five-year Wilkinson Microwave Anisotropy Probe observations: cosmological interpretation. *Astrophys. J.* **180** (suppl.), 330–376 (2009).
26. Malhotra, S. & Rhoads, J. E. Luminosity functions of Ly α emitters at redshifts $z=6.5$ and $z=5.7$: evidence against reionization at $z\leq 6.5$. *Astrophys. J.* **617**, L5–L8 (2004).
27. Becker, G. D., Rauch, M. & Sargent, W. L. W. The evolution of optical depth in the Ly α forest: evidence against reionization at $z\sim 6$. *Astrophys. J.* **662**, 72–93 (2007).
28. Willingale, R. *et al.* Testing the standard fireball model of gamma-ray bursts using late X-ray afterglows measured by Swift. *Astrophys. J.* **662**, 1093–1110 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Ph. Yock, B. Allen, P. Kubanek, M. Jelinek and S. Guziy for their assistance with the BOOTES-3 YA telescope observations (Supplementary Information). This work was partly based on observations obtained at the Gemini Observatory, which is operated by the Association of Universities for Research in Astronomy, Inc., under a cooperative agreement with the US National Science Foundation on behalf of the Gemini partnership: the National Science Foundation (United States), the Science and Technology Facilities Council (United Kingdom), the National Research Council (Canada), CONICYT (Chile), the Australian Research Council (Australia), the Ministério da Ciência e Tecnologia (Brazil) and SECYT (Argentina). This work was also partly based on observations made using ESO telescopes at the La Silla or Paranal observatories by G. Carraro, L. Schmidtbreich, G. Marconi, J. Smoker, V. Ivanov, E. Mason and M. Huertas-Company. The UKIRT is operated by the Joint Astronomy Centre on behalf of the UK Science and Technology Facilities Council. R.J.F. acknowledges a Clay Fellowship.

Author Contributions Triggering observations: N.R.T., D.B.F., A.J.L., E.B., J.S.B., D.P., J. Greiner, A.J.C.-T., A.d.U.P.; analysis of ground-based data: N.R.T., D.B.F., A.J.L., E.B., K.W., J.P.U.F., A.C., J.S.B., J.F., J.D., J. Gorosabel, B.C., D.P., J.R.M., T. Krühler, A.J.C.-T., A.d.U.P., C.G.M.; Swift analysis: P.A.E., R.L.C.S., K.P., R.W., A.J.L., N.R.T., N.G., D.W., P.S., T.S.; observations at various observatories and their automation to accept GRB overrides: A.J.A., A.A., T. Kerr, T.N., A.W.S., K.R., T.W. All authors made contributions through their involvement in the programmes from which the data derive, and contributed to the interpretation, content and discussion presented here. Writing was led by N.R.T., A.J.L., D.B.F. and E.B.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to N.R.T. (nrt3@star.le.ac.uk).

LETTERS

GRB 090423 at a redshift of $z \approx 8.1$

R. Salvaterra¹, M. Della Valle^{2,3,4}, S. Campana¹, G. Chincarini^{1,5}, S. Covino¹, P. D'Avanzo^{1,5}, A. Fernández-Soto⁶, C. Guidorzi⁷, F. Mannucci⁸, R. Margutti^{1,5}, C. C. Thöne¹, L. A. Antonelli⁹, S. D. Barthelmy¹⁰, M. De Pasquale¹¹, V. D'Elia⁹, F. Fiore⁹, D. Fugazza¹, L. K. Hunt⁸, E. Maiorano¹², S. Marinoni^{13,14}, F. E. Marshall¹⁰, E. Molinari^{1,13}, J. Nousek¹⁵, E. Pian^{16,17}, J. L. Racusin¹⁵, L. Stella⁹, L. Amati¹², G. Andreuzzi¹³, G. Cusumano¹⁸, E. E. Fenimore¹⁹, P. Ferrero²⁰, P. Giommi²¹, D. Guetta⁹, S. T. Holland^{10,22,23}, K. Hurley²⁴, G. L. Israel⁹, J. Mao¹, C. B. Markwardt^{10,23,25}, N. Masetti¹², C. Pagani¹⁵, E. Palazzi¹², D. M. Palmer¹⁸, S. Piranomonte⁹, G. Tagliaferri¹ & V. Testa⁹

Gamma-ray bursts (GRBs) are produced by rare types of massive stellar explosion. Their rapidly fading afterglows are often bright enough at optical wavelengths that they are detectable at cosmological distances. Hitherto, the highest known redshift for a GRB was $z = 6.7$ (ref. 1), for GRB 080913, and for a galaxy was $z = 6.96$ (ref. 2). Here we report observations of GRB 090423 and the near-infrared spectroscopic measurement of its redshift, $z = 8.1^{+0.1}_{-0.3}$. This burst happened when the Universe was only about 4 per cent of its current age³. Its properties are similar to those of GRBs observed at low/intermediate redshifts, suggesting that the mechanisms and progenitors that gave rise to this burst about 600,000,000 years after the Big Bang are not markedly different from those producing GRBs about 10,000,000,000 years later.

GRB 090423 was detected by NASA's Swift satellite on 23 April 2009 at 07:55:19 UT as a double-peaked burst of duration $T_{90} = 10.3 \pm 1.1$ s. As observed by Swift's Burst Alert Telescope (BAT)⁴, it had a 15–150-keV fluence of $F = (5.9 \pm 0.4) \times 10^{-7}$ erg cm⁻² and a peak energy of $E_p = 48^{+6}_{-5}$ keV (errors at the 90% confidence level). Its X-ray afterglow was identified by Swift's X-ray Telescope (XRT), which began observations 73 s after the BAT trigger⁵. A prominent flare was detected at $t \approx 170$ s in the X-ray light curve, which shows a typical 'steep decay/plateau/normal decay' behaviour (Fig. 1). Swift's Ultraviolet/Optical Telescope did not detect a counterpart even though it started making settled exposures only 77 s after the trigger⁶. A 2- μ m counterpart was detected with the United Kingdom Infra-Red Telescope, Hawaii, 20 min after the trigger⁷. Evidence that this burst occurred at high redshift was given by the Gamma-Ray Burst Optical/Near-Infrared Detector (GROND, Chile) multiband imager (from the g' band to the K band), which indicated a photometric redshift of $z = 8.0^{+0.4}_{-0.8}$ (ref. 7).

We used the 3.6-m Telescopio Nazionale Galileo (TNG, La Palma) with the Near Infrared Camera Spectrometer (NICS) and the Amici prism to obtain a low-resolution ($R \approx 50$) spectrum of GRB 090423 ~ 14 h after the trigger. NICS/Amici is an ideal instrument to detect spectral breaks in the continuum of faint objects because of its high efficiency and wide simultaneous spectral coverage (0.8–2.4 μ m).

The spectrum (Fig. 2) reveals a clear break at a wavelength of 1.1 μ m (ref. 8). We derive a spectroscopic redshift for the GRB of $z = 8.1^{+0.1}_{-0.3}$ (ref. 9; see Supplementary Information, section 3), interpreting the break as Lyman- α absorption in the intergalactic medium. No other significant absorption features were detected. This result is consistent, within the errors, with the measurement reported in ref. 7.

At $z \approx 8.1$, GRB 090423 has a prompt-emission rest-frame duration of only $T_{90,rf} = 1.13 \pm 0.12$ s in the redshifted 15–150-keV energy band, an isotropic equivalent energy of $E_{iso} = (1.0 \pm 0.3) \times 10^{53}$ erg in the redshifted 8–1,000-keV energy band¹⁰ and a peak energy of $E_{p,rf} = 437 \pm 55$ keV. The short duration and the high peak energy are consistent both with the distribution of long bursts, linked to massive stellar collapse, and with the population of short bursts, thought to arise from the merger of binary compact stars^{11,12}. Although the analysis of the spectral lag between the high- and low-energy channels in the BAT band is inconclusive about the classification of GRB 090423, the high E_{iso} argues in favour of a long GRB. The fact that GRB 090423 matches the $E_{iso}-E_{p,rf}$ correlation of long GRBs within 0.5σ further supports this classification¹³ (Supplementary Fig. 2).

The rest-frame γ -ray and X-ray light curves of GRB 090423 are remarkably akin to those of long GRBs at low, intermediate and high redshifts (Fig. 1), suggesting similar physics and interaction with the circumburst medium. The near-infrared light curve of GRB 090423 ~ 15 h after the trigger shows a temporal decay with a power-law index of $\alpha_0 \approx 0.5$, which is markedly different from the decay observed at X-ray energies during the same time interval, which has a power-law index of $\alpha_{X,2} \approx 1.3$ (Supplementary Fig. 3 and Supplementary Information, section 2). As for other lower-redshift GRBs, this behaviour is difficult to reconcile with standard afterglow models, although the sampling of the near-infrared light curve is too sparse for any firm conclusion to be drawn.

The spectral energy distribution of near-infrared afterglow is well fitted by a power law with an index of $\beta = 0.4^{+0.2}_{-1.4}$ and an equivalent interstellar extinction of $E(B-V) < 0.15$, assuming dust reddening

¹INAF, Osservatorio Astronomico di Brera, Via E. Bianchi 46, 23807 Merate (LC), Italy. ²INAF, Osservatorio Astronomico di Capodimonte, Salita Moiraniello 16, 80131 Napoli, Italy. ³European Southern Observatory, 85748 Garching, Germany. ⁴International Centre for Relativistic Astrophysics, Piazzale della Repubblica 2, 65122 Pescara, Italy. ⁵Dipartimento di Fisica G. Occhialini, Università di Milano Bicocca, Piazza della Scienza 3, 20126 Milano, Italy. ⁶Instituto de Física de Cantabria, CSIC-Universidad Cantabria, Avenida de los Castros s/n, 39005 Santander, Spain. ⁷Dipartimento di Fisica, Università di Ferrara, Via Saragat 1, 44100 Ferrara, Italy. ⁸INAF, Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy. ⁹INAF, Osservatorio Astronomico di Roma, Via di Frascati 33, 00040 Monte Porzio Catone, Rome, Italy. ¹⁰NASA, Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ¹¹Mullard Space Science Laboratory (UCL), Holmbury Road, Holmbury St Mary, Dorking RH5 6NT, UK. ¹²INAF, IASF di Bologna, Via Gobetti 101, 40129 Bologna, Italy. ¹³INAF, Fundación Galileo Galilei, Rambla José Ana Fernández Pérez 7, 38712 Breña Baja, TF - Spain. ¹⁴Università degli Studi di Bologna, Via Ranzani 1, 40127 Bologna, Italy. ¹⁵Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, Pennsylvania 16802, USA. ¹⁶INAF, Trieste Astronomical Observatory, Via G. B. Tiepolo 11, 34143 Trieste, Italy. ¹⁷Scuola Normale Superiore, Piazza dei Cavalieri 1, 56100 Pisa, Italy. ¹⁸INAF, Istituto di Astrofisica Spaziale e Fisica Cosmica di Palermo, Via Ugo La Malfa 153, 90146 Palermo, Italy. ¹⁹Los Alamos National Laboratory, PO Box 1663, Los Alamos, New Mexico 87545, USA. ²⁰Thüringer Landessternwarte Tautenburg, Sternwarte 5, 07778 Tautenburg, Germany. ²¹ASI Science Data Center, ASDC c/o ESRIN, Via G. Galilei, 00044 Frascati, Italy. ²²Universities Space Research Association, 10211 Wincopin Circle, Suite 500, Columbia, Maryland 21044, USA. ²³Centre for Research and Exploration in Space Science and Technology, Code 668.8, Greenbelt, Maryland 20771, USA. ²⁴Space Sciences Laboratory, 7 Gauss Way, University of California, Berkeley, California 94720-7450, USA. ²⁵Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA.

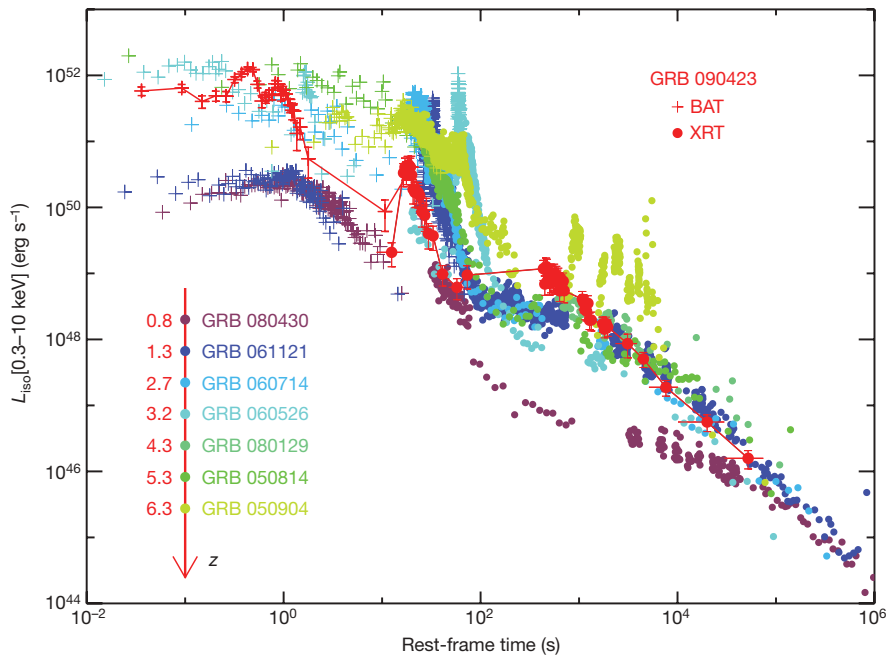


Figure 1 | Rest-frame γ -ray and X-ray light curves for bursts at different redshifts. BAT and XRT light curves of GRB 090423 (red data) in the source rest frame. Errors in luminosity, L_{iso} , are at the 1σ level; horizontal bars refer to the integration time interval. The XRT 0.3–10-keV light curve shows a prominent flare at a rest-frame time of $t_{\text{rf}} \approx 18$ s (also detected by BAT), and a flat phase (with a power-law index of $\alpha_{\text{X},1} = 0.13 \pm 0.11$) followed by a typical decay with a power-law index of $\alpha_{\text{X},2} = 1.3 \pm 0.1$. We compare the light curves of GRB 090423 with those of seven GRBs in the redshift interval

consistent with the Small Magellanic Cloud⁹. On the other hand, the analysis of the XRT data in the time interval 3,900–21,568 s suggests the presence of intrinsic absorption (in excess of the Galactic value) with an equivalent hydrogen column density of $N_{\text{H}}(z) = 6.8^{+5.6}_{-5.3} \times 10^{22} \text{ cm}^{-2}$ (90% confidence level; Supplementary Information, section 1). The low value of the dust extinction coupled with a relatively high value of N_{H} suggests that GRB 090423 originates from a region with low dust content relative to those of low- z GRBs¹⁴, but one similar to that of the high- z GRB 050904, for which $z = 6.3$ (ref. 15). Because the absorbing medium must be thin from the point of view of ‘Thomson’ scattering, the metallicity of the circumburst medium can be constrained to be $>4\%$ of the solar value, Z_{\odot} . The implication is that previous supernova explosions have already enriched the host galaxy of GRB 090423 to more than the critical metallicity, $Z \approx 10^{-4} Z_{\odot}$ (ref. 16), that prevents the formation of very massive stars (population III stars). Therefore, the progenitor of GRB 090423 should belong to a second stellar generation. Its explosion injected fresh metals into the interstellar medium, further contributing to the enrichment of its host galaxy. Its existence empirically supports the cosmological models^{17,18} in which stars and galaxies, already enriched in metals, are in place only $\sim 600,000,000$ yr after the Big Bang. Long GRBs are mostly associated with star-forming dwarf galaxies, which are thought to be the dominant population of galaxies in the early Universe¹⁹. The fact that GRB 090423 appears to have exploded in an environment similar to that of low- z GRB hosts²⁰ is in agreement with this.

The occurrence of a GRB at $z \approx 8$ has important implications for the cosmic history of these objects^{21–24}. In a first, simple, approach, we can assume that GRBs trace the cosmic star formation history, given the well-known link between long GRBs and the deaths of massive stars²⁵, and that GRBs are well described by a universal luminosity function. However, under these assumptions the expected number of bursts at $z \geq 8$ with an observed photon peak flux larger than or equal to that of GRB 090423 is extremely low: $\sim 4 \times 10^{-4}$ in ~ 4 yr of Swift operation (Supplementary Fig. 6 and Supplementary Information, section 4). Hence, one or both of the above assumptions may be an

oversimplification^{24,26}. The detection of a very high- z burst such as GRB 090423 could be accommodated if the GRB luminosity function were shifted towards higher luminosities according to $(1+z)^{\delta}$ with

0.8–6.3. The bursts are selected from among those showing a canonical three-phase behaviour (steep decay/plateau/normal decay) in the X-ray light curve and without a spectral break between BAT and XRT, allowing the spectral calibration of the BAT signal into the 0.3–10-keV energy band. The light curves of GRB 090423 do not have any distinguishing features relative to those of the lower-redshift bursts, suggesting that the physical mechanism that causes the GRB and its interaction with the circumburst medium are similar at $z \approx 8.1$ and at lower redshifts.

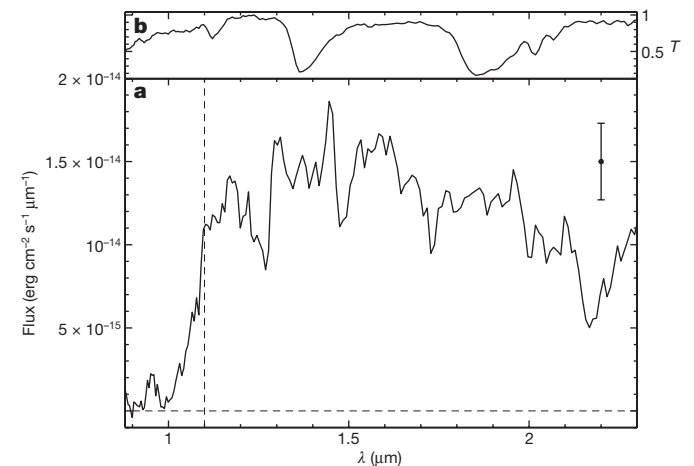


Figure 2 | TNG spectrum of the near-infrared afterglow. **a**, Spectrum of GRB 090423 obtained using the Amici prism on the TNG. The sharp break at wavelength $\lambda \approx 1.1 \mu\text{m}$, which is due to H I absorption in the intergalactic medium at the wavelength of the Lyman- α line, implies that $z = 8.1^{+0.1}_{-0.3}$. The spectrum has been smoothed with a boxcar filter of width $\Delta = 25$ pixels (where one pixel corresponds to $\sim 0.006 \mu\text{m}$ at $\lambda = 1.1 \mu\text{m}$). The absolute flux calibration was obtained by matching the almost simultaneous GROND photometric measurements⁷. The wavelength calibration was obtained from the TNG archive and adjusted to the wavelengths of the main atmospheric bands. The error bar corresponds to 1σ uncertainty as measured on the smoothed spectrum. The confidence level of the Lyman- α break detection is $\geq 4\sigma$. See also Supplementary Information, section 3. **b**, Plot of transmittance, T (the atmospheric transparency convolved with the instrumental response). The system has a significant sensitivity down to $0.9 \mu\text{m}$, and no instrumental or atmospheric effect could explain the abrupt flux break observed in the spectrum of GRB 090423.

$\delta \gtrsim 1.5$, or if the GRB formation rate were strongly enhanced in galaxies with $Z \lesssim 0.2Z_{\odot}$. The requirement for evolution may be mitigated if we assume a very high star formation rate at $z > 8$. However, we note that the need for evolution is strongly supported by both the large number of Swift detections at $z > 2.5$ (ref. 24) and the number of bursts with peak luminosities in excess of $10^{53} \text{ erg s}^{-1}$ (ref. 26). A possible explanation is that high-redshift galaxies are characterized by a top-heavy (bottom-light) stellar initial mass function with a higher incidence of massive stars than in the local Universe²⁷, providing an enhanced number of GRB progenitors. Such objects could be the main agents responsible for completing the reionization of the Universe^{19,28–30}.

Received 3 June; accepted 19 August 2009.

- Greiner, J. *et al.* GRB 080913 at redshift 6.7. *Astrophys. J.* **693**, 1610–1620 (2009).
- Iye, M. *et al.* A galaxy at a redshift $z = 6.96$. *Nature* **443**, 186–188 (2006).
- Komatsu, E. *et al.* Five-year Wilkinson Microwave Anisotropy Probe observations: cosmological interpretation. *Astrophys. J. Suppl. Ser.* **180**, 330–376 (2009).
- Palmer, D. M. *et al.* GRB 090423: Swift-BAT refined analysis. *GCN Circ.* **9204** (2009).
- Stratta, G. & Perri, M. GRB 090423: Swift-XRT refined analysis. *GCN Circ.* **9212** (2009).
- De Pasquale, M. & Krimm, H. GRB090423 - Swift/UVOT upper limits. *GCN Circ.* **9210** (2009).
- Tanvir, N. *et al.* A γ -ray burst at a redshift of $z \approx 8.2$. *Nature* doi:10.1038/nature08459 (this issue).
- Thoene, C. C. *et al.* GRB 090423: TNG Amici spectrum. *GCN Circ.* **9216** (2009).
- Fernández-Soto, A. *et al.* GRB 090423: refined TNG analysis. *GCN Circ.* **9222** (2009).
- von Kienlin, A. GRB 090423: Fermi GBM observation. *GCN Circ.* **9229** (2009).
- Mészáros, P. Gamma-ray bursts. *Rep. Prog. Phys.* **69**, 2259–2322 (2006).
- Zhang, B. Gamma-ray bursts in the Swift era. *Chin. J. Astron. Astrophys.* **7**, 1–50 (2007).
- Amati, L. *et al.* On the consistency of peculiar GRBs 060218 and 060614 with the $E_{p,i} - E_{iso}$ correlation. *Astron. Astrophys.* **463**, 913–919 (2007).
- Schady, P. *et al.* Dust and gas in the local environments of gamma-ray bursts. *Mon. Not. R. Astron. Soc.* **377**, 273–284 (2007).
- Stratta, G. *et al.* Dust properties at $z = 6.3$ in the host galaxy of GRB 050904. *Astrophys. J.* **661**, 9–12 (2007).
- Schneider, R. *et al.* First stars, very massive black holes, and metals. *Astrophys. J.* **571**, 30–39 (2002).
- Springel, V. *et al.* Simulations of the formation, evolution and clustering of galaxies and quasars. *Nature* **435**, 629–636 (2005).
- Nagamine, K. *et al.* Tracing early structure formation with massive starburst galaxies and their implications for reionization. *N. Astron.* **50**, 29–34 (2006).
- Choudhury, T. R., Ferrara, A. & Gallerani, S. On the minimum mass of reionization sources. *Mon. Not. R. Astron. Soc.* **385**, L58–L62 (2008).
- Fruchter, A. S. *et al.* Long γ -ray bursts and core-collapse supernovae have different environments. *Nature* **7092**, 463–468 (2006).
- Lamb, D. Q. & Reichart, D. E. Gamma-ray bursts as a probe of the very high redshift universe. *Astrophys. J.* **536**, 1–18 (2000).
- Guetta, D., Piran, T. & Waxman, E. The luminosity and angular distributions of long-duration gamma-ray bursts. *Astrophys. J.* **619**, 412–419 (2005).
- Bromm, V. & Loeb, A. High-redshift gamma-ray bursts from population III progenitors. *Astrophys. J.* **642**, 382–388 (2006).
- Salvaterra, R. & Chincarini, G. The gamma-ray burst luminosity function in the light of the Swift 2 year data. *Astrophys. J.* **656**, 49–52 (2007).
- Woosley, S. E. & Bloom, J. S. The supernova gamma-ray burst connection. *Annu. Rev. Astron. Astrophys.* **44**, 507–556 (2006).
- Salvaterra, R. *et al.* Evidence for luminosity evolution of long gamma-ray bursts in Swift data. *Mon. Not. R. Astron. Soc.* **396**, 299–303 (2009).
- Chary, R.-R. The stellar initial mass function at the epoch of reionization. *Astrophys. J.* **680**, 32–40 (2008).
- Bolton, J. S. & Haehnelt, M. G. The observed ionization rate of the intergalactic medium and the ionizing emissivity at $z \geq 5$: evidence for a photon-starved and extended epoch of reionization. *Mon. Not. R. Astron. Soc.* **382**, 325–341 (2007).
- Furlanetto, S. R. & Mesinger, A. The ionizing background at the end of reionization. *Mon. Not. R. Astron. Soc.* **394**, 1667–1673 (2009).
- Stiavelli, M. *From First Light to Reionization: The End of the Dark Ages* (Wiley-VCH, 2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the TNG staff for useful support during target-of-opportunity observations, in particular A. Fiorenzano, N. Sacchi and A. G. de Gurtubai Escudero. We thank A. Ferrara for discussions. This research was supported by the Agenzia Spaziale Italiana, the Ministero dell'Università e della Ricerca, the Ministero degli Affari Esteri, NASA and the US National Science Foundation.

Author Contributions Direct analysis of the Swift data: S. Campana, G. Chincarini, C.G., R.M., S.D.B., M.D.P., F.E.M., J.N., J.L.R., G. Cusumano, E.E.F., P.G., S.T.H., J.M., C.B.M., C.P., D.M.P.; analysis of the TNG and photometric data: M.D.V., S. Covino, P.D'A., A.F.-S., C.C.T., L.A.A., F.M., V.D'E., F.F., D.F., L.K.H., E. Maiorano, E. Molinari, S.M.; management of optical follow-up: P.D'A., L.A.A., V.D'E., E. Maiorano, S.M., G.A., P.F., G.L.I., N.M., E.P., S.P., G.T., V.T.; interpretation of the GRB properties: R.S., M.D.V., S. Campana, G. Chincarini, S. Covino, P.D'A., A.F.-S., C.G., R.M., C.C.T., L.A., E.P., L.S., K.H.; modelling of the GRB luminosity function: R.S., M.D.V., S. Campana, G. Chincarini, C.G., D.G., G.T. All authors made contributions through their involvement in the programmes from which the data derive, and contributed to the interpretation, content and discussion presented here.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.S. (salvaterra@mib.infn.it).

Acceleration of neutral atoms in strong short-pulse laser fields

U. Eichmann^{1,2}, T. Nubbemeyer¹, H. Rottke¹ & W. Sandner^{1,2}

A charged particle exposed to an oscillating electric field experiences a force proportional to the cycle-averaged intensity gradient. This so-called ponderomotive force¹ plays a major part in a variety of physical situations such as Paul traps^{2,3} for charged particles, electron diffraction in strong (standing) laser fields^{4–6} (the Kapitza–Dirac effect) and laser-based particle acceleration^{7–9}. Comparably weak forces on neutral atoms in inhomogeneous light fields may arise from the dynamical polarization of an atom^{10–12}; these are physically similar to the cycle-averaged forces. Here we observe previously unconsidered extremely strong kinematic forces on neutral atoms in short-pulse laser fields. We identify the ponderomotive force on electrons as the driving mechanism, leading to ultrastrong acceleration of neutral atoms with a magnitude as high as $\sim 10^{14}$ times the Earth's gravitational acceleration, g . To our knowledge, this is by far the highest observed acceleration on neutral atoms in external fields and may lead to new applications in both fundamental and applied physics.

The investigation has become possible through two recent findings concerning atomic ionization dynamics in strong laser fields. First, neutral atoms can survive a strong laser field in a (long-lived) excited state¹³, in which they can be detected directly in an atomic beam by means of a standard electron or ion detector¹⁴. Thus, any momentum transferred to the neutral atom can easily be detected. Second, according to the physical picture behind the excitation process, the excited electron behaves as a quasi-free electron during the laser pulse. More precisely, the excitation process can be viewed as a frustrated tunnel ionization¹⁴ within the three-step model for strong-field ionization¹⁵.

In the first step, the electron tunnels in the close vicinity of the maximum electric field of a laser cycle. The liberated electron is then driven by the laser field with an amplitude that slowly decreases with decreasing pulse intensity; in this way an active damping of the electronic motion takes place. After the laser pulse the electron is left with a drift energy too low to overcome the Coulomb potential of the ion and is recaptured into a Rydberg state. The quivering quasi-free electron experiences the ponderomotive force during the laser pulse owing to the intensity gradient in the focused laser beam. We will show here that the quiver motion of the electron is partially converted into centre-of-mass motion of the neutral atom, leading to a substantial acceleration. This results in a measurable momentum transfer to the atom despite the short interaction time in the femtosecond range. Remarkably, the ponderomotive effect is typically estimated to be negligible for these conditions^{16,17} with, however, a few exceptions¹⁸. We note that the investigation relies on the highly selective process of excitation of neutrals in a strong laser field, where kinematic effects are imparted only through the gradient of the laser field.

In the experiment we excite neutral He atoms in an effusive atomic beam using a perpendicularly intersecting focused laser beam. Using

the detection technique (see the Methods) we measure the distribution of excited He atoms on a detector as shown in Fig. 1. If, during the laser pulse, no momentum is transferred to the atoms, we would expect a slightly enlarged projected image of the (laser-intensity-dependent) distribution of excited atoms in the laser beam on the detector, that is, a distribution that extends along the laser beam direction (z axis), typically within the Rayleigh length, but with a very narrow radial distribution (r_D axis) of the order of the size of the laser beam waist.

In Fig. 1a, however, we see a strikingly large radial distribution of excited atoms with a strong maximum in the laser focal plane ($z = 0$) that obviously stems from a deflecting radial force during the laser pulse. In Fig. 1b the cut along the z axis (black curve) shows two maxima at roughly half the laser peak intensity $I_0/2$, where the net production rate of excited helium atoms He* is apparently maximum, whereas the He* signal at I_0 shows a pronounced minimum. However, the loss of neutral excited atoms is largely due to their radial deflection. The full projection (red dashed curve) shows only a slight decrease in signal, indicating that even at the highest intensities He atoms are excited. The data are taken at a low beam target pressure of $\simeq 5 \times 10^{-7}$ mbar. The radial deflection is unchanged when we increase the target pressure by more than a factor of 30. This excludes many-particle effects based on atom density or space charge as an origin of our observations. Furthermore, we emphasize that the radial distribution is unaltered whether the linear polarization of the laser beam is in the direction of the atomic beam or perpendicular to it. In this respect the intensity-dependent force very much resembles the ponderomotive force acting on charged particles. The question arises whether we can conclude that the ponderomotive force is responsible for the observed centre-of-mass motion of the neutral particle.

To shed light on the underlying process we first recall that the ponderomotive force F_p on a charged particle is given by (all equations are in atomic units):

$$F_p = -\frac{q^2}{4m\omega^2} \nabla |E_0|^2 \quad (1)$$

Here, m and q are the mass and the charge of the particle, respectively, $E(\mathbf{r}, t) = E_0(\mathbf{r}, t)\exp(i\omega t)$ is the electric field, ω is the field frequency and $E_0(\mathbf{r}, t)$ is the slowly varying field amplitude. Hence, in view of our frustrated tunnel ionization model, both the ionic core and the electron experience a mass-dependent ponderomotive force during the laser pulse. As a consequence of the mass dependency, however, the ionic core remains practically unaffected while the electron experiences a non-negligible ponderomotive force. This, in turn, means that to the first approximation a ponderomotive force acts directly on the centre-of-mass motion of the atom and leaves the recapturing process unaffected. This can be shown more rigorously: we derive the centre-of-mass motion from the Lorentz force (see the

¹Max-Born-Institute, Max-Born-Strasse 2a, 12489 Berlin, Germany. ²Institut für Optik und Atomare Physik, Technische Universität Berlin, 10632 Berlin, Germany.

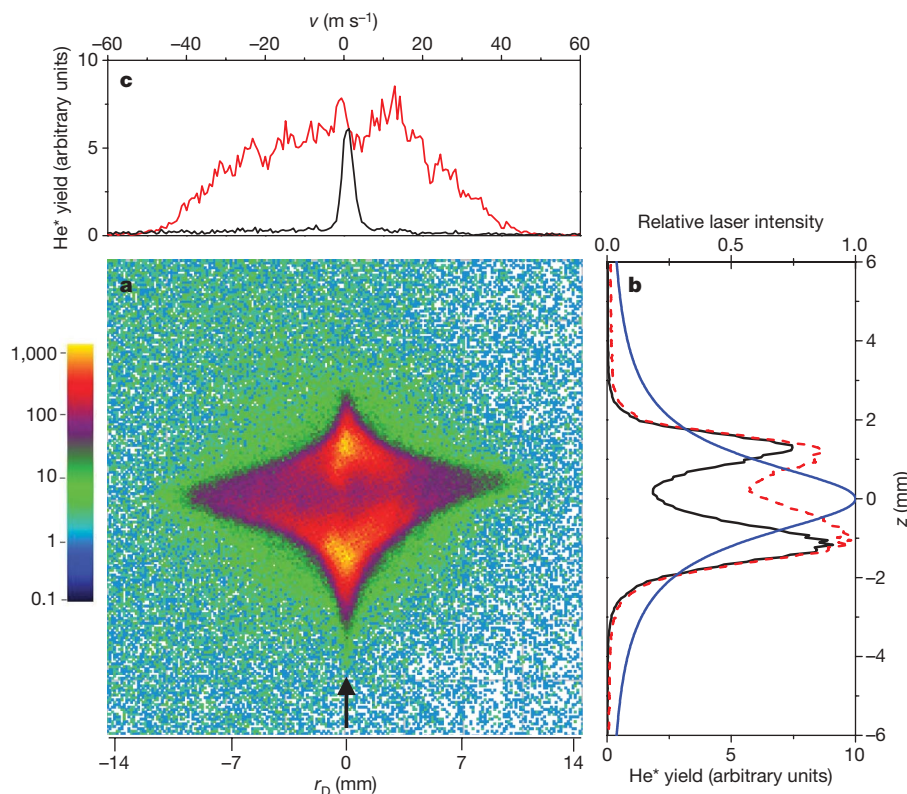


Figure 1 | Deflection of neutral He atoms after interaction with a focused laser beam. **a**, Distribution of excited He* atoms on the detector (colour scale, in number of atoms). The laser beam direction is indicated by the arrow. **b**, Cut through the atom distribution along the laser beam axis (z axis) at $r_D = 0$ mm (black curve) and full projection on z axis (dashed red curve) and intensity along the z axis in units of the laser peak intensity

$I_0 = 6.9 \times 10^{15} \text{ W cm}^{-2}$ (blue curve). **c**, Cuts through the distribution at $z = 0$ mm (red curve) and $z = -2.7$ mm (black curve). The black curve shows the velocity distribution of excited neutral atoms at a position unaffected by the ponderomotive force, showing essentially the ‘natural’ velocity spread, while the red curve shows the velocity gain through the ponderomotive force.

online-only Methods). Recalling the key point of our investigation, that the electron remains bound after interaction with the laser pulse, we are able to observe the centre-of-mass motion. Furthermore, by solving the coupled Lorentz equations for the electron and the ion, including the Coulomb potential, we can directly reproduce the capture process into bound Rydberg orbits and the force on the centre of mass.

According to our model we can rewrite equation (1) for the centre-of-mass position \mathbf{R} of the neutral atom:

$$M\ddot{\mathbf{R}}(t) = -\frac{1}{4m_e\omega^2}\nabla|\mathbf{E}_0|^2 \quad (2)$$

Here, M and m_e ($= 1$ atomic unit) are the masses of the atom and the electron, respectively, and $\ddot{\mathbf{R}}(t)$ is the second derivative of the centre-of-mass position \mathbf{R} with respect to time. To calculate the ponderomotive force explicitly, we assume a linearly polarized laser beam with a Gaussian spatial intensity distribution, which reads, in cylindrical coordinates:

$$I(\mathbf{r}) = |\mathbf{E}_0(\mathbf{r})|^2 = I_0 \left(1 + \left(\frac{z}{z_0} \right)^2 \right)^{-1} \exp \left(-\frac{2r^2}{r_0^2} \right) \quad (3)$$

where $r_0 = w_0 \sqrt{1 + (z/z_0)^2}$, w_0 is the beam waist. Evaluating the gradient in equation (2) with the intensity distribution given by equation (3), we obtain, for the radial component of the centre-of-mass position perpendicular to the laser beam direction:

$$\ddot{r}(t) = \frac{I(\mathbf{R})}{M\omega^2} \frac{r(t)}{r_0^2} f(t) \quad (4)$$

where $f(t)$ is the laser pulse envelope, which we assume to be of the form $f(t) = \exp(-t^2/\tau^2)$, where τ is the pulse width. From equation (4)

we find that the maximum force along the radial direction scales as r_0^{-1} . Similarly, one can show that it scales as z_0^{-1} along the laser beam direction. Because the Rayleigh length z_0 is typically a factor of 100 larger than the beam waist r_0 , the gradient and thus the ponderomotive force in the laser beam direction is much smaller than in the radial direction and can be neglected. (However, the situation would be very different if we used a short-pulse standing-wave laser field. We would then obtain a strong periodic intensity gradient on the scale of the laser wavelength, and might expect to see the Kapitza–Dirac effect for neutral atoms in an intense standing-wave laser field instead of electrons¹⁹).

To solve equation (4), we assume that the neutral atom does not move significantly during the laser pulse. Hence, we set $r(t) = r$ on the right-hand side of the equation, which allows us to solve equation (4) analytically for any initial position of an atom in the laser beam. We will concentrate our analysis on atoms located at the half beam size $r_0/2$, which experience the maximum force. Solving equation (4) for these conditions by integrating over the full laser pulse, we find the maximum velocity $v_{\max}(z)$:

$$v_{\max}(z) = \frac{I_0}{2M\omega^2 w_0} \frac{\sqrt{\pi} \exp(-0.5)}{\sqrt{1 + \left(\frac{z}{z_0} \right)^2}} \tau \quad (5)$$

If we evaluate equation (5) at the focal plane for He atoms exposed to our focused laser beam at maximum intensity, we obtain a velocity of about 55 m s^{-1} from which accelerations of about $2 \times 10^{14} \text{ g}$ can be deduced.

This exceeds the typical acceleration (deceleration) of neutral atoms²⁰ or molecules in external fields^{21,22}. Compared to laser-cooling experiments in a continuous-wave laser field, for instance, which are

based on photon momentum transfer with a typical deceleration of about $10^6 g$ (ref. 20), the acceleration present in our experiment is eight orders of magnitude larger. It is, to the best of our knowledge, by far the largest acceleration of neutral matter by electromagnetic fields ever observed.

Finally, we mention that a different but equivalent description of our process might be given in terms of an atom in the Kramers–Henneberger reference frame²³, which is expected to exhibit stable configurations in a strong electromagnetic field. Our observations of accelerated neutral atoms seem to be a direct confirmation of the existence of this exotic type of stable atom.

In Fig. 2 we show the results of a systematic investigation of He atoms exposed to laser light of different laser pulse lengths and laser pulse energy E_L , such that the laser intensity is kept constant. We obtain neutral excited atom distributions similar to the one shown in Fig. 1. The maximum radial deflection of the atoms along the laser beam axis, which can be converted into maximum velocity $v_{\max}(z)$, can reliably be extracted from the experimental data to compare with equation (5). Fitting equation (5) to the data gives very good agreement, confirming the validity of our model. The fit contains only three free parameters: a small velocity offset, a common constant

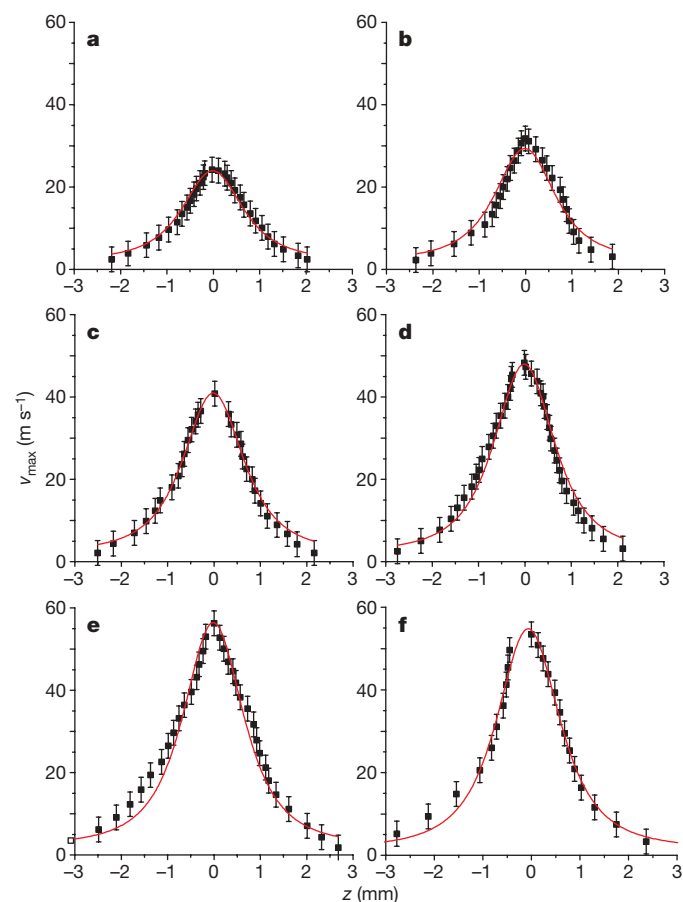


Figure 2 | Maximum velocity $v_{\max}(z)$ gained by neutral He atoms. **a–e**, $v_{\max}(z)$ extracted from measurements at constant laser peak intensity $I_0 = 2.8 \times 10^{15} \text{ W cm}^{-2}$ ($3.4 \times 10^{15} \text{ W cm}^{-2}$), but for different pulse energies E_L and pulse durations τ_{FWHM} : **a**, 600 μJ , 40 fs; **b**, 900 μJ , 60 fs; **c**, 1.2 mJ, 80 fs; **d**, 1.5 mJ, 100 fs; **e**, 1.8 mJ, 120 fs. FWHM, full-width at half-maximum. The red curves are fits to the data based on equation (5). The fitted common beam waist $w_0 = 16.0 \mu\text{m}$ is in good agreement with an independent experimental determination of $w_0 = 17.5 \pm 1.5 \mu\text{m}$ using the knife-edge method. It results, however, in a higher laser peak intensity, which is indicated in parentheses. **f**, $v_{\max}(z)$ for laser parameters $E_L = 1.8 \text{ mJ}$ and $\tau_{\text{FWHM}} = 40 \text{ fs}$ corresponding to $I_0 = 8.3 \times 10^{15} \text{ W cm}^{-2}$ ($10 \times 10^{15} \text{ W cm}^{-2}$). The error bars are estimated from the accuracy of the maximum deflection we determined.

beam waist w_0 , and a scaling factor for the absolute velocity. The best choice of the latter implies that the acceleration must last over the whole pulse duration, confirming the assumption that the electrons are set free early during the laser pulse.

Even then the observed velocities lie slightly above the theoretical prediction (see Fig. 3), which we attribute to absolute laser intensity uncertainties or a slightly non-Gaussian intensity distribution. Hence, we note that the radial velocity $v_{\max}(z)$ might serve as a valuable beam shape and intensity parameter along the laser beam axis. On the one hand it measures the peak intensity if the radial intensity distribution is known (for example, Gaussian), or, on the other hand, indicates systematic deviations from the Gaussian beam shape if the peak intensity is known (for example, from ionization experiments). Such information is otherwise rather difficult to obtain.

We have also measured the radial deflection of a beam of Ne atoms, which is, however, smaller than in He, as expected owing to the larger mass. In Fig. 3 we show the maximum velocity $v_{\max}(z=0)$ transferred to Ne atoms exposed to a laser beam kept at a constant laser intensity of $I_0 = 2.8 \times 10^{15} \text{ W cm}^{-2}$, but at different pulse durations and energies. For comparison, the maximum velocities for He are also shown. The data clearly show the excellent quantitative agreement of the observed deflection with our theoretical predictions. Within the range of our achievable pulse durations the deflection is proportional to the mass and the time the force is acting on the atom.

In conclusion, we have demonstrated ultrastrong acceleration of neutral atoms, using femtosecond laser pulses with intensities of up to $I_0 = 10^{16} \text{ W cm}^{-2}$. We present a quantitative theoretical model based on the concept that a bound electron temporarily undergoes a quasi-free oscillatory quiver motion during the laser pulse while still being coupled to the ion by the Coulomb force. The inhomogeneous field of a focused laser beam causes a ponderomotive force on the electron, resulting in a centre-of-mass acceleration of the whole atom. Many new applications may be envisioned, considering that the mechanism transfers the momenta of a large number (presently of the order of 10^3) of photons quasi-instantaneously onto neutral atoms, without any problems due to Doppler detuning, as in the case of resonant continuous-wave laser absorption. Atom optics, controlled atom deposition or controlled chemical reactions are just some of them. The acceleration process may be considerably refined by sophisticated spectral and temporal shaping of the accelerating laser beam, including the use of standing (or slowly moving) waves with much steeper field gradients, or the use of longer pulse durations for achieving substantially larger atomic momenta.

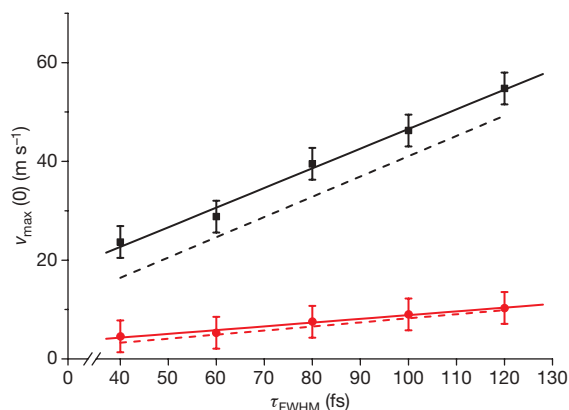


Figure 3 | Maximum velocity $v_{\max}(0)$ transferred to He and Ne at the focal plane as a function of the laser pulse duration at constant laser intensity. $v_{\max}(0)$ is plotted for He (black squares) and Ne (red circles) measured at $I_0 = 2.8 \times 10^{15} \text{ W cm}^{-2}$ ($3.4 \times 10^{15} \text{ W cm}^{-2}$); see explanation in Fig. 2. The dashed black and red curves show the calculated velocities, respectively, using equation (5) with the fitted beam waist, but omitting the scaling factor. The full curves are fits to the data. The error bars are determined as in Fig. 2.

METHODS SUMMARY

The observation of acceleration of neutral atoms in strong laser fields is based on the method of producing and detecting excited neutral atoms in strong laser fields introduced in ref. 14. In the online-only Methods we give a detailed description of how to adapt the method to the present experiments. Furthermore, we detail the theoretical background of the underlying process and derive the respective formulas used to describe our data.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 May; accepted 27 August 2009.

1. Kibble, T. W. B. Refraction of electron beams by intense electromagnetic waves. *Phys. Rev. Lett.* **16**, 1054–1056 (1966).
2. Boot, H. A. H., & Harvie, R. B. R.-S. Charged particles in a non-uniform radio-frequency field. *Nature* **180**, 1187 (1957).
3. Dehmelt, H. G. Radio-frequency spectroscopy of stored ions. *Adv. At. Mol. Phys.* **3**, 53–72 (1967).
4. Kapitza, P. & Dirac, P. The reflection of electrons from standing light waves. *Proc. Camb. Philos. Soc.* **29**, 297–300 (1933).
5. Bucksbaum, P. H., Schumacher, D. W. & Bashkansky, M. High-intensity Kapitza-Dirac effect. *Phys. Rev. Lett.* **61**, 1182–1185 (1988).
6. Freimund, D. L., Aflatooni, K. & Batelaan, H. Observation of the Kapitza-Dirac effect. *Nature* **413**, 142–143 (2001).
7. Tajima, T. & Dawson, J. M. Laser electron accelerator. *Phys. Rev. Lett.* **43**, 267–270 (1979).
8. Geddes, C. *et al.* High-quality electron beams from a laser wakefield accelerator using plasma-channel guiding. *Nature* **431**, 538–541 (2004).
9. Mourou, G., Tajima, T. & Bulanov, S. Optics in the relativistic regime. *Rev. Mod. Phys.* **78**, 309–371 (2006).
10. Gould, P. L., Ruff, G. A. & Pritchard, D. E. Diffraction of atoms by light: The near-resonant Kapitza-Dirac effect. *Phys. Rev. Lett.* **56**, 827–830 (1986).
11. Chu, S., Bjorkholm, J. E., Ashkin, A. & Cable, A. Experimental observation of optically trapped atoms. *Phys. Rev. Lett.* **57**, 314–317 (1986).
12. Grimm, R., Weidemüller, M. & Ovchinnikov, Y. Optical dipole traps for neutral atoms. *Adv. At. Mol. Phys.* **42**, 95–170 (2000).
13. de Boer, M. P. & Muller, H. G. Observation of large populations in excited states after short-pulse multiphoton ionization. *Phys. Rev. Lett.* **68**, 2747–2750 (1992).
14. Nubbemeyer, T., Gorling, K., Saenz, A., Eichmann, U. & Sandner, W. Strong-field tunneling without ionization. *Phys. Rev. Lett.* **101**, 233001 (2008).
15. Corkum, P. B. Plasma perspective on strong field multiphoton ionization. *Phys. Rev. Lett.* **71**, 1994–1997 (1993).
16. Krapchev, V. B. Kinetic theory of the ponderomotive effects in a plasma. *Phys. Rev. Lett.* **42**, 497–500 (1979).
17. McNaught, S. J., Knauer, J. P. & Meyerhofer, D. D. Photoelectron initial conditions for tunneling ionization in a linearly polarized laser. *Phys. Rev. A* **58**, 1399–1411 (1998).
18. Wells, E., Ben-Itzhak, I. & Jones, R. R. Ionization of atoms by the spatial gradient of the ponderomotive potential in a focused laser beam. *Phys. Rev. Lett.* **93**, 023001 (2004).
19. Batelaan, H. Illuminating the Kapitza-Dirac effect with electron matter optics. *Rev. Mod. Phys.* **79**, 929–941 (2007).
20. Chu, S. Nobel lecture: The manipulation of neutral particles. *Rev. Mod. Phys.* **70**, 685–706 (1998).
21. Stapelfeldt, H., Sakai, H., Constant, E. & Corkum, P. B. Deflection of neutral molecules using the nonresonant dipole force. *Phys. Rev. Lett.* **79**, 2787–2790 (1997).
22. Fulton, R., Bishop, A. I. & Barker, P. F. Optical Stark decelerator for molecules. *Phys. Rev. Lett.* **93**, 243004 (2004).
23. Henneberger, W. C. Perturbation method for atoms in intense light beams. *Phys. Rev. Lett.* **21**, 838–841 (1968).

Acknowledgements We thank F. Noack for technical support on the laser system and W. Becker, P. B. Corkum, H. R. Reiss and O. Smirnova for discussions.

Author Contributions U.E. and T.N. designed and performed the experiments and analysed the data. All authors contributed to the theoretical understanding and were involved in the completion of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to U.E. (ulli.eichmann@mbi-berlin.de).

METHODS

Experimental set-up and detection of excited neutral atoms. We use a slightly modified version of an experimental set-up that has been described elsewhere¹⁴. Briefly, a linearly polarized Ti:sapphire laser beam with a repetition rate of 500–1,000 Hz, a variable pulse width of $\tau_{\text{FWHM}} = 40\text{--}120$ fs and pulse energies up to 2.5 mJ is steered into a vacuum chamber with a base pressure of $\sim 10^{-9}$ mbar. It is focused by means of a plano-convex lens with a focal length of 0.2 m into an effusive beam of He or Ne atoms at right angles, where, among other processes, it excites neutral atoms. The beam target pressure is typically $\leq 5 \times 10^{-7}$ mbar. 0.38 m downstream in the direction of the atomic beam, a position-sensitive microchannel plate detector is located. The detector can be configured to detect either ions or electrons in the counting mode. Both configurations are suitable for excited neutral species detection, if they have an excitation energy of more than ~ 5 eV. In contrast to charged particles, which are easily accelerated by small electric fields (either deliberately applied or stemming from spurious charges) and hit the detector after a few microseconds, neutral atoms travel at thermal velocities reaching the detector only after hundreds of microseconds. Consequently, the excited state needs to survive long enough that the atom can reach the detector in an excited state. It has been found in earlier work¹⁴ that an intense laser interacting with a beam of rare gas atoms produces mainly Rydberg states with principal quantum numbers around $n = 10$, which decay partially to a metastable state with high excitation energy, for example, ~ 20 eV in He. The metastable state lives long enough to reach the detector. We estimate that about 1% of all excited atoms reach the detector in an excited state.

Owing to the divergence of the atomic beam, the signal on the detector in laser beam direction is an image enlarged by a factor of 2.3 of the distribution of excited atoms in the laser beam, in case the momentum transferred to the atom during the excitation can be neglected. This needs to be taken into account when transforming the detector signal along the z_0 coordinate into the laser beam z coordinate. The high detection sensitivity to accelerating forces during the laser pulse lies in the nature of neutral excited atoms travelling with thermal velocities. The time-of-flight distribution of He atoms at the microchannel plate detector peaks around 230 μs . Because momentum is only transferred to the atoms during the laser pulse duration, the starting point of the atoms is well defined. By setting a 40- μs -wide gate on the maximum of the time-of-flight distribution, a precise determination of the radial velocity can be calculated from the radial beam deflection on the detector. We note that the centre-of-mass motion of a neutral atom is largely unaffected by spurious electric fields or space charge.

Derivation of the ponderomotive centre-of-mass force on neutral atoms. To derive the centre-of-mass force on the neutral atom we solve the coupled Lorentz equations for the ion with mass m_1 and charge q_1 and electron with mass m_2 and charge q_2 including the Coulomb interaction between them. The two Lorentz equations in atomic units read:

$$\mathbf{F}_1 = m_1 \ddot{\mathbf{r}}_1 = q_1 (\mathbf{E}_1 + \mathbf{v}_1 \times \mathbf{B}_1) + \mathbf{F}_c \quad (6)$$

$$\mathbf{F}_2 = m_2 \ddot{\mathbf{r}}_2 = q_2 (\mathbf{E}_2 + \mathbf{v}_2 \times \mathbf{B}_2) - \mathbf{F}_c \quad (7)$$

where $\mathbf{E}_i \equiv \mathbf{E}(\mathbf{r}_i, t)$ and $\mathbf{B}_i \equiv \mathbf{B}(\mathbf{r}_i, t)$. The Coulomb force \mathbf{F}_c is given by $\mathbf{F}_c = q_1 q_2 \frac{(\mathbf{r}_1 - \mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|^3}$. Using the first-order electric and magnetic fields for a focused Gaussian beam²⁴, we can write the equations explicitly:

$$\begin{aligned} m_1 \ddot{x}_1 &= E_{1x} + \dot{y}_1 B_{1z} - \dot{z}_1 B_{1y} + F_{cx} \\ m_1 \ddot{y}_1 &= -\dot{x}_1 B_{1z} + F_{cy} \\ m_1 \ddot{z}_1 &= \dot{x}_1 B_{1y} + F_{cz} + E_{1z} \end{aligned} \quad (8)$$

$$\begin{aligned} m_2 \ddot{x}_2 &= -E_{2x} - \dot{y}_2 B_{2z} + \dot{z}_2 B_{2y} - F_{cx} \\ m_2 \ddot{y}_2 &= \dot{x}_2 B_{2z} - F_{cy} \\ m_2 \ddot{z}_2 &= -\dot{x}_2 B_{2y} - F_{cz} - E_{2z} \end{aligned} \quad (9)$$

Equation (2) can be derived analytically from the Lorentz forces using centre-of-mass and relative coordinates. We obtain the force on the centre-of-mass by using the following approximations. The fields are approximated to first order and we assume the slowly varying amplitude approximation; the relative velocity and position are approximated through the motion of the electron in the electric field in polarization (x) direction, neglecting the Coulomb force.

By solving equations (8) and (9) numerically without the aforementioned approximations, we obtain, for specific initial conditions, the bound trajectories of the tunnel electron at the end of the laser pulse as described in ref. 14. Most importantly, by evaluating the centre-of-mass velocity from the numerical calculations we confirm the results of equation (2).

To give an interpretation of why the electron remains bound, we first note that its motion during the laser pulse is mainly driven by the laser field. If we suppose that the electron starts at a maximum of a field cycle close to the nucleus, that is, where the Coulomb field is strong, the electron then revisits its starting (inner turning) point periodically with the laser frequency. Here, the attractive force is strongly peaked. For example, if the electron is set free at a distance of 10 atomic units (a.u.) from the nucleus, the Coulomb force is 10^{-2} a.u. acting for roughly a tenth of the laser cycle duration. If one compares this periodical force with the steady ponderomotive force F_p , which is of the order of 10^{-4} a.u. for $I = 0.1$ a.u. and beam waist of $w_0 = 16 \mu\text{m}$ we find an attractive net force.

Sign of the centre-of-mass force on a neutral atom in a laser field. A focus of a continuous-wave laser beam can act as a quasi-static dipole trap for neutral atoms¹² provided the photon energy of the laser is much below the lowest excited state. In this case the atom is a strong-field seeker. In our experiment, the force on the centre-of-mass motion according to equation (4) is outward bound, that is, the atom in the strong short-pulsed laser field is a low-field seeker, although the photon energy of the laser field is much below the lowest excited state. Obviously, the effective polarizability of an atom in a strong laser field, where the electron can be treated as quasi-free, differs in sign from a ground-state atom in a weak continuous laser field, where the strongly bound electron is only weakly perturbed by the laser field. According to the analysis of ref. 25 there is a relationship between the dynamical polarizability of an atom and the ponderomotive potential for a free electron U_p , from which the ponderomotive force can be obtained by $F_p = -\nabla U_p$. For a quasi-free electron, as is the case for our experiments, $U_p = \frac{e^2}{4m\omega^2} E^2$. The more general ponderomotive potential for a bound electron with a binding frequency ω_b is given by $U_p = \frac{e^2}{4m(\omega^2 - \omega_b^2)} E^2$, which can be derived from the Lorentz model of atomic polarizability. It is immediately obvious that for a strongly bound electron with $\omega \ll \omega_b$ the sign of the ponderomotive potential is opposite and the absolute value is much smaller than for a free electron.

24. Quesnel, B. & Mora, P. Theory and simulation of the interaction of ultraintense laser pulses with electrons in vacuum. *Phys. Rev. E* **58**, 3719–3732 (1998).

25. Eberly, J. H., Javanainen, J. & Rzazewski, K. Above-threshold ionization. *Phys. Rep.* **204**, 331–383 (1991).

Preserving electron spin coherence in solids by optimal dynamical decoupling

Jiangfeng Du¹, Xing Rong¹, Nan Zhao², Ya Wang¹, Jiahui Yang¹ & R. B. Liu²

To exploit the quantum coherence of electron spins in solids in future technologies such as quantum computing^{1,2}, it is first vital to overcome the problem of spin decoherence due to their coupling to the noisy environment. Dynamical decoupling^{3–9}, which uses stroboscopic spin flips to give an average coupling to the environment that is effectively zero, is a particularly promising strategy for combating decoherence because it can be naturally integrated with other desired functionalities, such as quantum gates. Errors are inevitably introduced in each spin flip, so it is desirable to minimize the number of control pulses used to realize dynamical decoupling having a given level of precision. Such optimal dynamical decoupling sequences have recently been explored^{9–12}. The experimental realization of optimal dynamical decoupling in solid-state systems, however, remains elusive. Here we use pulsed electron paramagnetic resonance to demonstrate experimentally optimal dynamical decoupling for preserving electron spin coherence in irradiated malonic acid crystals at temperatures from 50 K to room temperature. Using a seven-pulse optimal dynamical decoupling sequence, we prolonged the spin coherence time to about 30 μ s; it would otherwise be about 0.04 μ s without control or 6.2 μ s under one-pulse control. By comparing experiments with microscopic theories, we have identified the relevant electron spin decoherence mechanisms in the solid. Optimal dynamical decoupling may be applied to other solid-state systems, such as diamonds with nitrogen-vacancy centres^{13–15}, and so lay the foundation for quantum coherence control of spins in solids at room temperature.

The idea of dynamical decoupling originated in spin echo^{16,17} in nuclear magnetic resonance, in which a single flip of nuclear spins by a radio pulse makes the noises ‘felt’ by the spins before and after the pulse cancel each other, at least partially. Recently, progress has been made on electron spin echo in solid-state systems, such as quantum dots and impurities^{18–20}. The standard periodic Carr–Purcell–Meiboom–Gill (CPMG) pulse sequences¹⁷ can keep the spin coherence longer by decoupling a spin from its environment²¹. Iteratively concatenated sequences^{5–8} can realize the dynamical decoupling to an arbitrarily high order of precision, but the number of pulses increases exponentially with the controlling order. A remarkable advance in dynamical decoupling theory is the discovery by Uhrig⁹ of optimal sequences, which realize the N th order of dynamical decoupling precision with the minimum number (N) of spin-flip pulses. The Uhrig dynamical decoupling (UDD) was first discovered for a specific spin-boson model⁹, and was later conjectured¹⁰ and rigorously proved¹¹ to be model-independent. The first experimental study of UDD was done with trapped ions with the noise mimicked by modulating the controlling system¹². Here we report the experimental demonstration of UDD for electron spins in solids, namely, radical electron spins in γ -irradiated malonic acid single crystals, from 50 K to room temperature. Our results demonstrate that the UDD performs better than the periodic CPMG dynamical

decoupling (PDD). The experimental data from different samples under various conditions are in good agreement with calculations based on microscopic theories (with only one fitting parameter, fixed at the same value for all samples under all conditions). We demonstrate the strengths as well as some limitations of the dynamical decoupling in forestalling various decoherence mechanisms relevant in the solids.

A general spin–environment interaction is described by a Hamiltonian as

$$\mathcal{H} = \gamma_e (\mathbf{B}_0 + \delta\mathbf{B}) \cdot \boldsymbol{\sigma} / 2 + \mathcal{H}_E \quad (1)$$

where γ_e is the electron gyromagnetic ratio, $\boldsymbol{\sigma}$ indicates the Pauli matrices ($\sigma_x, \sigma_y, \sigma_z$) representing the quantized spin moment, \mathbf{B}_0 is the external magnetic field, $\delta\mathbf{B}$ is the fluctuation part of the local magnetic field due to coupling to the environment, and \mathcal{H}_E contains many-body interactions within the environment. $\delta\mathbf{B}$ contains a static fluctuation resulting from thermal distribution of environmental states, usually referred to as inhomogeneous broadening. Also, a dynamical quantum fluctuation will be built up, since an environmental state with a certain eigenvalue of $\delta\mathbf{B}$, driven by the many-body interaction \mathcal{H}_E , will evolve into a superposition of different local-field eigenstates. The spin precesses about the local magnetic field $\mathbf{B}_0 + \delta\mathbf{B}$, and the fluctuation $\delta\mathbf{B}$ leads to the decay of the spin polarization $\langle \sigma \rangle$ that quantifies the quantum coherence.

Dynamical decoupling aims to reduce the effective random field by flipping the spin from time to time. For example, by applying instantaneous π -rotations at t_1, t_2, \dots, t_k along the directions $\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_k$, respectively, the time-averaged effective Hamiltonian \mathcal{H}_{eff} is given by $e^{-i\mathcal{H}_{\text{eff}}t} \equiv (-i)^k e^{-i\mathcal{H}(t-t_k)} \sigma_k \dots \sigma_2 e^{-i\mathcal{H}(t_2-t_1)} \sigma_1 e^{-i\mathcal{H}t_1}$ with $\sigma_i \equiv \mathbf{n}_i \cdot \boldsymbol{\sigma}$ which flips the spin along \mathbf{n}_i . The target is to eliminate the fluctuating field to a given order of t , that is, $\mathcal{H}_{\text{eff}} = \mathcal{H}_E + O(t^{N+1})$. As errors are inherently introduced by the controlling pulses, we should ideally use the minimum number of pulses for a given order of decoupling precision. For a pure dephasing model in which the random field is fixed along a certain direction, say, the z axis, the UDD contains N spin-flip pulses applied at

$$t_j = t \sin^2 \frac{j\pi}{2(N+1)}, \quad j = 1, 2, \dots, N \quad (2)$$

to eliminate the coupling up to $O(t^{N+1})$ (see Fig. 1c). For comparison, the PDD has the periodic timing $t_j = t(2j-1)/(2N)$. Below we abbreviate the N -pulse UDD and PDD as UDD N and PDD N , respectively. Note that UDD1 and UDD2 are identical to the standard Hahn echo (PDD1) and Carr–Purcell control (PDD2)¹⁷, respectively. But the higher order Uhrig sequences become non-periodic.

To test the dynamical decoupling for spins in solids, we took γ -irradiated malonic acid single crystals as the benchmark systems. A schematic of a unit cell of the crystal is shown in Fig. 1a. The γ -irradiation may remove one hydrogen atom from a methylene group ($-\text{CH}_2$) and

¹Hefei National Laboratory for Physical Sciences at Microscale and Department of Modern Physics, University of Science and Technology of China, Hefei, Anhui 230026, China.

²Department of Physics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China.

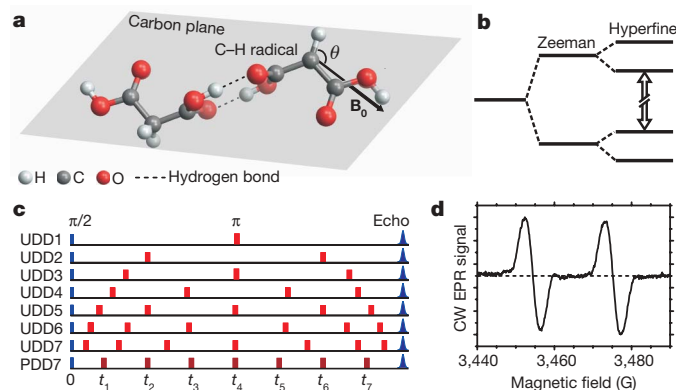


Figure 1 | System and methods for the dynamical decoupling experiments. **a**, A unit cell of a malonic acid crystal contains two $(\text{CH}_2)(\text{COOH})_2$ molecules. The C–H bond of a radiation-generated radical and the magnetic field \mathbf{B}_0 are in the carbon plane and form an angle $\theta \approx 2\pi/3$. **b**, Energy levels of the electron and proton spins in a radical. The arrow indicates the transition used in the pulsed EPR. **c**, Pulse sequences for UDD1–7 and PDD7. **d**, First-derivative CW EPR spectrum measured at 9.722 GHz.

create a radical ($-\dot{\text{C}}\text{H}$). The spin-1/2 of the unpaired carbon valence electron is the object under consideration (referred to as qubit hereafter for clarification). We estimated the radical concentrations of the three single crystals used in the experiments at 32, 8 and 0.6 p.p.m. (see Supplementary Information for methods). Individually accessing single qubits is not necessary for testing dynamical decoupling, and we instead addressed ensembles of electron spins for enhanced signals. An additional advantage of using spin ensembles is that dynamical decoupling can be tested on unwanted couplings between qubits, an issue pertinent to large-scale quantum computing. Figure 1d shows a typical electron spin resonance in the continuous wave (CW) electron paramagnetic resonance (EPR) spectrum. The resonance is split owing to the hyperfine coupling (Fig. 1b) with the spin of the α -proton²², the hydrogen nucleus in the radical. The crystals are oriented such that the hyperfine interaction with the α -proton is $\mathcal{H}_{\text{C-H}} \approx 2\pi(A_0^z I_0^z + A_0^{xz} I_0^x) \sigma_z/2$, with $A_0^z = -45$ MHz, and $A_0^{xz} = 26$ MHz (see Supplementary Information for details). Here the electron spin flip terms associated with $\sigma_{x/y}$ have been dropped as the Zeeman energy $\gamma_e B_0$ is much greater than the hyperfine interaction strength A . Such coherent electron–nuclear spin coupling has been studied for the purpose of quantum information processing²³.

In the dynamical decoupling experiments, the spin precession was initiated from the thermal equilibrium by a $\pi/2$ pulse. Then N spin-flip pulses in the UDD or PDD timing were applied to protect the spin coherence. The areas of the echoes at t were measured as the spin coherence.

Echo signals in samples of various radical concentrations measured at various temperatures are shown in Figs 2 and 3, and compared with the theoretical calculation. Preservation of the spin coherence by dynamical decoupling is evidenced by the elongation of the coherence time with increasing the number of control pulses. The overall decoherence profile is non-exponential, with an initial exponential decay followed by a rapid super-exponential drop. To quantify the coherence preservation by dynamical decoupling control, we derived the decoherence time T_2 when the signal drops to $1/e$ its value at $t = 0$ which was obtained by extrapolation using exponential fitting of the initial decay (see Supplementary Information for details). The decoherence times in the 0.6 p.p.m. sample at 50 K under UDD1–7 and PDD1–7 control are listed in Table 1. It is clearly seen that UDD outperforms PDD in preserving the spin coherence.

Electron spin decoherence in solids can be caused by any of the following mechanisms: (1) direct relaxation due to spontaneous emission of photons; (2) spin–lattice relaxation via phonon scattering and spin–orbit interaction; (3) hyperfine interaction with nuclear spins; (4) coupling between qubits; and (5) coupling to other impurity

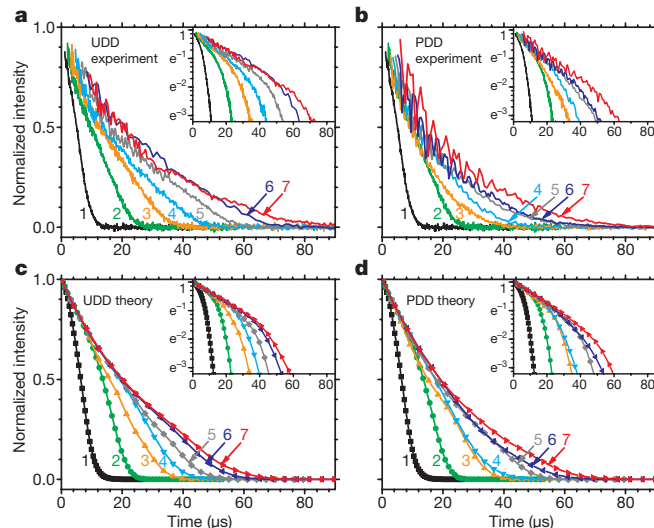


Figure 2 | Electron spin decoherence under UDD and PDD control. **a**, **b**, The UDD (**a**) and PDD (**b**) experiments. The integer associated with each curve indicates the number of control pulses. The sample had a radical concentration 0.6 p.p.m. and the temperature was 50 K. **c**, **d**, The theoretical calculations for **a** and **b**, respectively. Insets, semi-logarithmic plots of the same data; the variables plotted on the axes are the same as in the main panels.

electron spins. The direct photon emission rate in our resonator was estimated at $\sim 10^{-4} \text{ s}^{-1}$ (see Supplementary Information), so this particular mechanism can be disregarded given the timescale considered here. We clarified possible contributions of the other mechanisms using experiments for samples with various radical concentrations at various temperatures and theoretical calculations based on microscopic models.

Experiments at various temperatures (Fig. 3a) singled out the effect of spin–lattice relaxation. To obtain a sufficiently strong echo signal at room temperature, the 32 p.p.m. sample was used. We found the spin decoherence time (T_2 , in μs) to be 7.7 ± 0.4 , 7.7 ± 0.4 , 7.7 ± 0.3 and 7.6 ± 0.3 for temperature (T , in K) $T = 50$, 100, 200 and 300, respectively. In previous CPMG experiments²⁴, a thermal activation of spin–lattice relaxation around 150 K was observed. Here in our own experiments, any such thermal activation was too small to be resolved to the precision of the decoherence time (ΔT_2). Considering that the spin–lattice relaxation time measured at 50 K was as long as 940 μs (Supplementary Information), the spin decoherence rate attributable to phonon scattering at room temperature was below $\delta T_2/T_2^2 \approx 1/(150 \mu\text{s})$. Such a small phonon scattering effect is probably due to the weak spin–orbit coupling in carbon atoms. Thus we expect that the coherence preservation effect of dynamical decoupling shown in Fig. 2, which was measured at 50 K for a sufficiently strong signal in the low concentration sample, would be similar if measured at room temperature.

Hyperfine interaction with nuclear spins results in a random local magnetic field (Overhauser field), $\delta \hat{B}_{\text{hf}} = \gamma_e^{-1} 2\pi \sum_{i>0, \alpha=x,y,z} A_i^{\alpha} I_i^{\alpha}$, where A_i is the hyperfine coupling tensor to the i th nuclear spin \mathbf{I}_i . The thermal distribution of the Overhauser field (that is, inhomogeneous broadening) deduced from the resonance line width in the CW EPR spectrum (Fig. 1d) is $\Delta B_{\text{hf}} \approx 2.0$ G, consistent with the previous measurement²². The free-induction decay of the electron spin coherence due to the inhomogeneous broadening would have a timescale $T_2^* = \sqrt{2}(\gamma_e \Delta B_{\text{hf}})^{-1} \approx 40$ ns. Such rapid decay was not resolved in our pulsed EPR spectrometer. The effect of the static inhomogeneous broadening, however, is fully eliminated from the echo signals.

More relevant is the dynamical fluctuation of the local Overhauser field. The dynamical fluctuation $\delta \hat{B}_{\text{hf}}(t)$ can be driven by pair-wise nuclear-spin flip-flops due to the dipole–dipole interaction (see

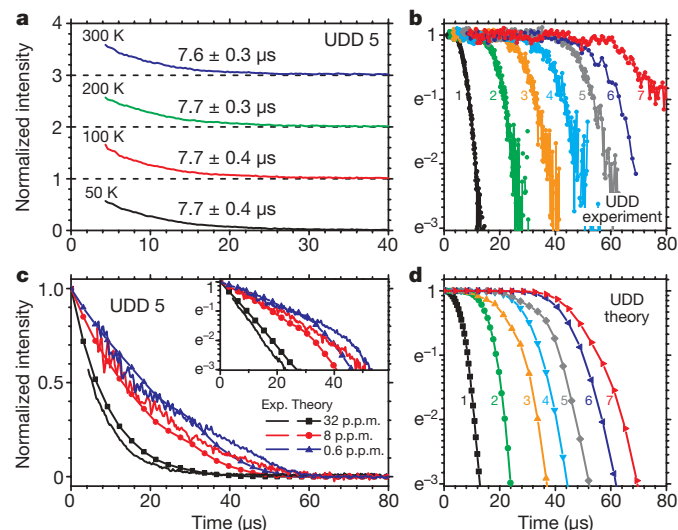


Figure 3 | Effects of various decoherence mechanisms in malonic acid crystals. **a**, UDD5 echo signals of the 32 p.p.m. sample at 50, 100, 200 and 300 K. Values of T_2 are shown next to each curve. **b**, UDD5 echo signals of the 32, 8 and 0.6 p.p.m. samples at 50 K. The curves with symbols are theoretical results. Inset, semi-logarithmic plot of the same data; the variables plotted on the axes are the same as in the main panel. **c**, UDD1–7 echo signals of the 0.6 p.p.m. sample at 50 K, each divided by the exponential factor which fits the initial stage of decoherence. **d**, UDD1–7 echo signals, calculated considering only the dynamical Overhauser field fluctuation driven by nuclear dipolar interaction.

Supplementary Information). The coupled electron–nuclear spins in the γ -irradiated malonic acid crystal provide an ideal platform for studying the paradigmatic problem of decoherence of a single qubit or centre spin embedded in an interacting spin bath^{25–28}. We calculated the electron spin decoherence in the nuclear spin baths using a quantum theory developed²⁹ to properly take into account the many-body correlations in the bath (Supplementary Information). The result is shown in Fig. 3d.

Dynamical Overhauser field fluctuation can also be caused by the random displacements of nuclei relative to the electron spin, especially in the case of nuclei near the radical. This effect is possible in the molecular crystals, which are loosely bound by hydrogen bonds and van der Waals forces. For example, at 4.2 K the radical would have two stable conformations separated by a low barrier, with the C–H bond tilted away from the carbon plane by $\pm 12^\circ$ (ref. 30). At slightly higher temperatures, the radical randomly hops between the two conformations and appears to be in the middle. This may result in extra dynamical fluctuation of the Overhauser field. Calculation of this effect, however, requires detailed knowledge of the electronic states, geometrical conformations, and hopping dynamics of radicals in the molecular crystal, which is not yet available.

Another component of the random local field comes from the dipolar interaction between a ‘qubit’ spin and other radical electron spins. This mechanism is particularly relevant for large-scale quantum computing with many qubits, where unwanted interaction between qubits is unavoidable. The static inhomogeneous broadening $\delta\hat{B}_{ee}$ due to the electron–electron spin interaction, proportional to the radical concentration, ranges roughly from 0.001 to 0.05 G in our samples, which is far less than the Overhauser field broadening and hence is unimportant in the CW EPR spectrum or the free-induction decay. But in contrast to nuclear spins, the ‘bath’ electron spins would be flipped together with the ‘qubit’ spin when their frequencies fall into the spectrum of the short control π -pulses. As a result, $\delta\hat{B}_{ee}$ was not fully eliminated and became important in the echo signals. The effect of the electron–electron spin interaction is shown in Fig. 3b. The decoherence time under UDD5 control in the 32 p.p.m. sample ($7.7 \pm 0.4 \mu\text{s}$) is much shorter than that in the

Table 1 | Decoherence times of electron spins under UDD and PDD control

No. of pulses	Decoherence time (μs)			
	UDD experiment*	UDD theory†	PDD experiment‡	PDD theory§
1	6.2 ± 0.1	7.56	6.2 ± 0.1	7.56
2	13.3 ± 0.4	15.7	13.3 ± 0.4	15.7
3	18.5 ± 0.5	20.1	13.5 ± 0.4	20.9
4	21.4 ± 0.4	24.5	17.6 ± 1.2	22.0
5	24.3 ± 0.9	25.7	18.8 ± 0.9	24.8
6	30.2 ± 1.5	28.1	20.2 ± 1.0	25.5
7	27.1 ± 0.9	29.0	21.9 ± 2.5	27.0

* For data shown in Fig. 2a.

† For data shown in Fig. 2c.

‡ For data shown in Fig. 2b.

§ For data shown in Fig. 2d.

0.6 p.p.m. sample ($24.3 \pm 0.9 \mu\text{s}$). This indicates that the electron–electron spin interaction dominates the decoherence in the high concentration sample. We confirmed this conclusion by calculation, considering the inhomogeneous broadening due to the electron spin interaction and the finite duration of the control pulses (Fig. 3b, see Supplementary Information for methods). Coupling to impurities should have little effect, since those impurities, if any, would be far off-resonant and hence any inhomogeneous broadening on their account would be totally removed by spin echoes.

All the experimental data in Figs 2a, 2b and 3b are in good agreement with the calculations (Figs 2c, 2d and 3b), which took into account the static inhomogeneous broadening due to electron–electron spin interaction and the dynamical Overhauser field fluctuation. The dynamical field fluctuation caused by random displacement of nuclei was described by a phenomenological decay factor $\exp(-t/T_2')$, with T_2' being the only fitting parameter and fixed the same ($30 \mu\text{s}$) in all the calculations. The decoherence due to random conformation-hopping may account for the saturation of the spin coherence time at about $30 \mu\text{s}$ in the 0.6 p.p.m. sample as the UDD goes beyond the sixth order (Fig. 2a). In samples with higher radical concentrations, the saturated coherence time is much shorter (for example, about $7.7 \mu\text{s}$ in the 32 p.p.m. sample). This is because dynamical decoupling can not decouple the qubits from the other electron spins in near-resonance. If we consider only the dynamical Overhauser field fluctuation driven by nuclear dipolar interaction, the calculation (Fig. 3d) shows that the decoherence under dynamical decoupling control presents a plateau with width increasing roughly linearly with the order of dynamical decoupling, and under UDD7 the coherence persists beyond $50 \mu\text{s}$. These features are indeed in agreement with the experimental data excluding the initial exponential decay (Fig. 3c). In the plateau regime, the decoherence would be dominated by the other mechanisms, such as the electron–electron interaction and the random conformation-hopping. This explains the initial exponential decay and the overall non-exponential profile seen in Figs 2 and 3. The small but noticeable discrepancy between the experimental and theoretical results may be ascribed to the lack of an accurate model of the hyperfine interaction, a precise estimation of radical concentrations, and a full understanding of the molecule conformation dynamics. Further study of the microscopic decoherence mechanisms in the solids is desirable.

METHODS SUMMARY

Malonic acid single crystals were grown from saturated aqueous solutions by slow evaporation at 5°C , γ -irradiated at room temperature with various doses, and aged at 67°C for 15 h to eliminate unstable radicals. The crystal structure was determined by X-ray diffraction. The radical concentrations were estimated from the areas of CW EPR resonances compared with a standard sample.

In the dynamical decoupling experiments, the spin precession was triggered by a $\pi/2$ -rotation using a 28-ns pulse. The spins were flipped by 56-ns pulses in the UDD or PDD timing. The area of the echo signal was measured. To avoid the interference of unwanted echoes in the multi-pulse experiments¹⁷, phase cycling was applied. The magnetic field and the pulse frequency were 3,477 G and 9.722 GHz for the 32 p.p.m. sample, 3,454 G and 9.664 GHz for the 8 p.p.m. sample, and 3,447 G and 9.645 GHz for the 0.6 p.p.m. sample, respectively. The signals near $t = 0$ were not measured owing to the dead window of the

spectrometer but obtained by extrapolation using exponential fitting of an initial range of data. The standard deviations of the fitting were added to the error bars of the decoherence times.

The photon emission rate was estimated considering the Purcell enhancement factor of our cavity. The nuclear spin bath dynamics were calculated using the cluster correlation expansion²⁹ including up to four-spin clusters. Including the nearest 500 nuclear spins around a radical was sufficient to produce converged results. The decoherence due to interaction with a 'bath' electron spin was calculated considering the rotation under the finite-duration pulses, the contributions from different electron spins were summed, and the results were averaged on position configurations.

Received 17 June; accepted 27 August 2009.

- Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
- Awschalom, D. D., Loss, D. & Samarth, N. (eds) *Semiconductor Spintronics and Quantum Computation* (Springer, 2002).
- Viola, L., Knill, E. & Lloyd, S. Dynamical decoupling of open quantum systems. *Phys. Rev. Lett.* **82**, 2417–2421 (1999).
- Kern, O. & Alber, G. Controlling quantum systems by embedded dynamical decoupling schemes. *Phys. Rev. Lett.* **95**, 250501 (2005).
- Khodjasteh, K. & Lidar, D. A. Fault-tolerant quantum dynamical decoupling. *Phys. Rev. Lett.* **95**, 180501 (2005).
- Santos, L. F. & Viola, L. Enhanced convergence and robust performance of randomized dynamical decoupling. *Phys. Rev. Lett.* **97**, 150501 (2006).
- Yao, W., Liu, R. B. & Sham, L. J. Restoring coherence lost to a slow interacting mesoscopic spin bath. *Phys. Rev. Lett.* **98**, 077602 (2007).
- Witzel, W. M. & Das Sarma, S. Concatenated dynamical decoupling in a solid-state spin bath. *Phys. Rev. B* **76**, 241303 (2007).
- Uhrig, G. S. Keeping a quantum bit alive by optimized π -pulse sequences. *Phys. Rev. Lett.* **98**, 100504 (2007).
- Lee, B., Witzel, W. M. & Das Sarma, S. Universal pulse sequence to minimize spin dephasing in the central spin decoherence problem. *Phys. Rev. Lett.* **100**, 160505 (2008).
- Yang, W. & Liu, R. B. Universality of Uhrig dynamical decoupling for suppressing qubit pure dephasing and relaxation. *Phys. Rev. Lett.* **101**, 180403 (2008).
- Biercuk, M. J. *et al.* Optimized dynamical decoupling in a model quantum memory. *Nature* **458**, 996–1000 (2009).
- Jelesko, F. *et al.* Observation of coherent oscillation of a single nuclear spin and realization of a two-qubit conditional quantum gate. *Phys. Rev. Lett.* **93**, 130501 (2004).
- Gaebel, T. *et al.* Room-temperature coherent coupling of single spins in diamond. *Nature Phys.* **2**, 408–413 (2006).
- Childress, L. *et al.* Coherent dynamics of coupled electron and nuclear spin qubits in diamond. *Science* **314**, 281–285 (2006).
- Hahn, E. Spin echoes. *Phys. Rev.* **80**, 580–594 (1950).
- Schweiger, A. & Jeschke, G. *Principles of Pulse Electron Paramagnetic Resonance* (Oxford Univ. Press, 2001).
- Petta, J. R. *et al.* Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309**, 2180–2184 (2005).
- Berezovsky, J., Mikkelsen, M. H., Stoltz, N. G., Coldren, L. A. & Awschalom, D. D. Picosecond coherent optical manipulation of a single electron spin in a quantum dot. *Science* **320**, 349–352 (2008).
- Tyryshkin, A. M., Lyon, S. A., Astashkin, A. V. & Raitsimring, A. M. Electron spin relaxation times of phosphorus donors in silicon. *Phys. Rev. B* **68**, 193207 (2003).
- Morton, J. J. L. *et al.* Bang-bang control of fullerene qubits using ultrafast phase gates. *Nature Phys.* **2**, 40–43 (2006).
- McConnell, H. M., Heller, C., Cole, T. & Fressenden, R. W. Radiation damage in organic crystals. I. $\text{CH}(\text{COOH})_2$ in malonic acid. *J. Am. Chem. Soc.* **82**, 766–775 (1960).
- Hodges, J. S., Yang, J. C., Remanathan, C. & Cory, D. G. Universal control of nuclear spins via anisotropic hyperfine interactions. *Phys. Rev. A* **78**, 010303(R) (2008).
- Dalton, L. R., Kwiram, A. L. & Cowen, J. A. Electron spin-lattice and cross relaxation in irradiated malonic acid. *Chem. Phys. Lett.* **14**, 77–81 (1972).
- Prokof'ev, N. V. & Stamp, P. C. E. Theory of the spin bath. *Rep. Prog. Phys.* **63**, 669–726 (2000).
- de Sousa, R. & Das Sarma, S. Theory of nuclear-induced spectral diffusion: spin decoherence of phosphorus donors in Si and GaAs quantum dots. *Phys. Rev. B* **68**, 115322 (2003).
- Witzel, W. M. & Das Sarma, S. Quantum theory for electron spin decoherence induced by nuclear spin dynamics in semiconductor quantum computer architectures: spectral diffusion of localized electron spins in the nuclear solid-state environment. *Phys. Rev. B* **74**, 035322 (2006).
- Yao, W., Liu, R. B. & Sham, L. J. Theory of electron spin decoherence by interacting nuclear spins in a quantum dot. *Phys. Rev. B* **74**, 195301 (2006).
- Yang, W. & Liu, R. B. Quantum many-body theory of qubit decoherence in a finite-size spin bath. *Phys. Rev. B* **78**, 085315 (2008).
- McCalley, R. C. & Kwiram, A. L. ENDOR studies at 4.2 K of the radicals in malonic acid single crystals. *J. Chem. Phys.* **97**, 2888–2903 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements J.D. thanks J. F. Chen for discussions on sample preparation. This work was supported by the National Natural Science Foundation of China, the Chinese Academy of Sciences, the Ministry of Education of PRC, the National Fundamental Research Program 2007CB925200, and Hong Kong GRF Projects CUHK401906 and CUHK402209.

Author Contributions J.D. conceived and designed the experiment; J.D., X.R. and Y.W. performed the EPR measurements; J.D. and J.Y. prepared the samples; R.B.L. and N.Z. formulated the theory; N.Z. performed the calculations; J.D. and R.B.L. analysed the experimental and theoretical data; and R.B.L. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.D. (djf@ustc.edu.cn) or R.B.L. (rblu@phy.cuhk.edu.hk).

A dearth of intermediate melts at subduction zone volcanoes and the petrogenesis of arc andesites

Olivier Reubi^{1†} & Jon Blundy¹

Andesites represent a large proportion of the magmas erupted at continental arc volcanoes and are regarded as a major component in the formation of continental crust¹. Andesite petrogenesis is therefore fundamental in terms of both volcanic hazard and differentiation of the Earth. Andesites typically contain a significant proportion of crystals showing disequilibrium petrographic characteristics indicative of mixing or mingling between silicic and mafic magmas, which fuels a long-standing debate regarding the significance of these processes in andesite petrogenesis² and ultimately questions the abundance of true liquids with andesitic composition. Central to this debate is the distinction between liquids (or melts) and magmas, mixtures of liquids with crystals, which may or may not be co-genetic. With this distinction comes the realization that bulk-rock chemical analyses of petrologically complex andesites can lead to a blurred picture of the fundamental processes behind arc magmatism. Here we present an alternative view of andesite petrogenesis, based on a review of quenched glassy melt inclusions trapped in phenocrysts, whole-rock chemistry, and high-pressure and high-temperature experiments. We argue that true liquids of intermediate composition (59 to 66 wt% SiO₂) are far less common in the sub-volcanic reservoirs of arc volcanoes than is suggested by the abundance of erupted magma within this compositional range. Effective mingling within upper crustal magmatic reservoirs obscures a compositional bimodality of melts ascending from the lower crust, and masks the fundamental role of silicic melts (≥ 66 wt% SiO₂) beneath intermediate arc volcanoes. This alternative view resolves several puzzling aspects of arc volcanism and provides important clues to the integration of plutonic and volcanic records.

Quenched glassy melt inclusions trapped in phenocrysts provide snapshots of liquid (melt) evolution during crystallization. As such, they represent the only unequivocal information on the composition of true volcanic liquids in petrologically complex magmas. A compilation of published melt inclusion data ($n = 2,582$) covering more than 85 arc volcanoes extracted from the GEOROC global compilation shows a range of SiO₂ content from 40 to 83 wt% (calculated on an H₂O-free basis). Although the range of SiO₂ content is continuous, the distribution is strongly bimodal, with major peaks around 54 and 76 wt% (Fig. 1a). There is a marked paucity of melt inclusions with intermediate (andesite) composition, particularly between 59 and 66 wt% SiO₂, despite the abundance of volcanic rocks of this composition in arcs. Minima at intermediate compositions are also observed for MgO, CaO, FeO and K₂O. A similar conclusion has been drawn³ from a compilation of melt inclusions for igneous rocks from all tectonic settings.

Melt inclusions show a good correlation with the bulk composition of their host rock, for host compositions <55 and >71 wt% SiO₂. The correlation persists for rocks up to 59 wt% SiO₂, although these samples occasionally also contain melt inclusions significantly

more evolved than their host rock. The vast majority of intermediate rocks contain melt inclusions that are rhyolitic in composition, irrespective of host-rock composition (Fig. 1c, d). The striking feature of Fig. 1 is that the pronounced intermediate composition peak observed in the bulk rock record is displaced towards significantly more silicic compositions in the melt inclusion record. This shift is independent of the mineral phase in which the melt inclusion is hosted (Fig. 1d) and cannot, therefore, result from post-entrapment crystallization of the host mineral or other systematic biases in the melt inclusion record (Methods).

It is widely held that melt inclusions in arc magmas record volatile-saturated conditions^{4,5}. H₂O and CO₂ contents in melt inclusions indicate typical pressures of entrapment below 550 and 350 MPa for mafic and silicic melts, respectively⁵, reflecting crystallization in the upper crust. Melt inclusions from individual volcanoes show steep to nearly sub-vertical negative correlations between H₂O and SiO₂ (Fig. 2), due to varying degrees of decompression-driven crystallization⁴. This process does not, however, appear to drive appreciable chemical differentiation of magmas owing to the inefficient separation of crystal and melt phases. Melt inclusions suggest that low-pressure crystallization of mafic melts can produce residual melts with ≤ 60 wt% SiO₂ but only rarely extends to high-silica andesite (Figs 1c and 2). This limit also corresponds to the most evolved composition of aphyric (crystal-free) andesites, for example^{6,7}, and to the most evolved compositions erupted from mafic cones and shield volcanoes in arcs⁸. The melt inclusion record suggests that this threshold corresponds to the upper limit of vapour-saturated differentiation of mafic magmas in the upper crust. Residual melts in crystal cumulates or groundmasses certainly evolve beyond 60 wt% SiO₂, but the high crystallinity involved precludes efficient separation of melt from crystals. In summary, the melt inclusion record indicates that melts reaching the upper crust, where they crystallize mostly in response to degassing, are compositionally bimodal with a gap from 60 to ≥ 65 wt% SiO₂. It is not exceptional for this gap to extend to 70 wt% SiO₂. Low-pressure crystallization of mafic melts partially fills the gap, but the proportion of unequivocal melts with 60–66 wt% SiO₂ remains remarkably low in shallow magmatic systems.

Experimental crystallization of hydrous basaltic melts over a range of crustal pressures^{9–14} elucidates the possible liquid lines of descent during crustal differentiation. Experimentally produced melts describe curved fractionation trends that culminate in the compositions of silicic magmas and melt inclusions (Fig. 3), demonstrating that fractionation of mafic magmas at some level within the crust is capable of producing these silicic liquids. A further contribution may come from crustal contamination, as often required by whole rock isotopic data^{9,15,16}. Both mafic and silicic melt inclusions show a good match to experimental liquids, suggesting that melt inclusions preserve true melt compositions produced by crystallization of hydrous basaltic

¹Department of Earth Sciences, University of Bristol, Wills Memorial Building, Bristol BS8 1RJ, UK. †Present address: Institute of Isotope Geochemistry and Mineral Resources, ETH Zurich, CH-8092 Zurich, Switzerland.

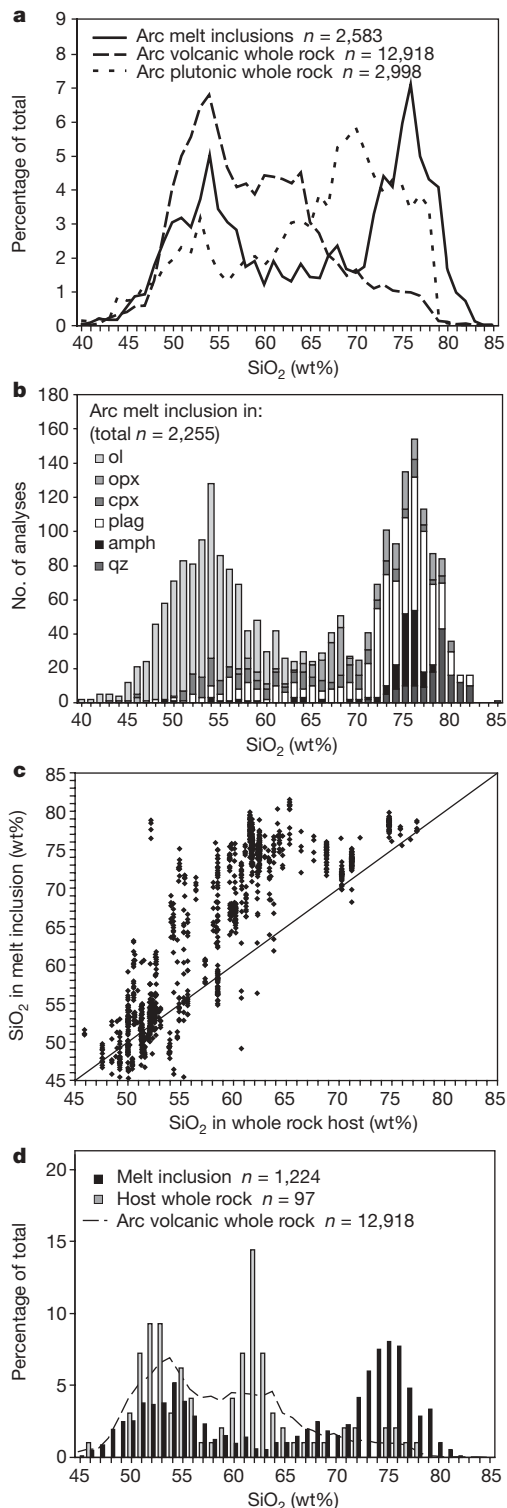


Figure 1 | SiO_2 contents (H_2O -free) of melt inclusions in arc magmas. **a**, Global compilations of arc magma melt inclusion and bulk rock compilation for North, Central and South America (GEOROC database). Plutonic curve shown for comparison is a compilation of data for the Lachlan fold belt, Australia (B. W. Chappell, personal communication), Adamello batholith, Italy (P. Ulmer, personal communication) and Sierra Nevada batholith, USA (NAVDAT compilation for Cretaceous plutonic rocks). **b**, Melt inclusion compositions in arc rocks, classified by host mineral phase. **c**, Compositions of melt inclusions versus host rock compositions for arc volcanoes (subset of the data in **a**). **d**, Distribution of melt inclusions and their host rock compositions for the data subset in **c**.

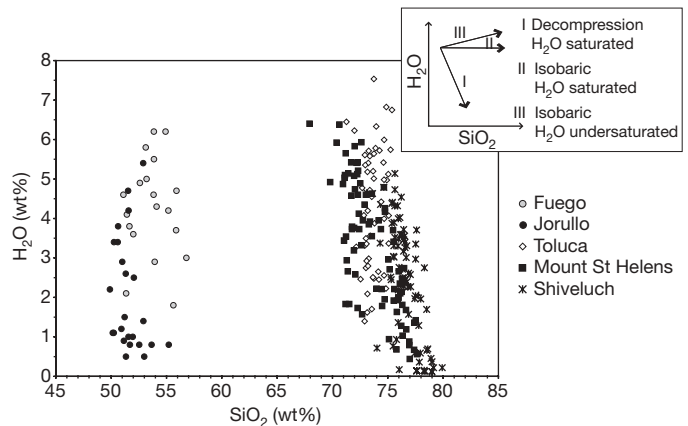


Figure 2 | H_2O - SiO_2 contents of melt inclusions from several arc volcanoes. SiO_2 contents are calculated on an H_2O -free basis, and chemical analyses are normalized to a total of 100%; data are taken from refs 4, 28–31, and Smith, V. C. *et al.*, manuscript in preparation. Inset, schematic trends anticipated for different crystallization mechanisms. Trends for individual volcanoes are indicative of various degrees of decompression-crystallization in vapour-saturated melts, and indicate that relatively limited differentiation occurs in the upper crust.

melts. However, only a limited number of intermediate melt inclusions correspond to experimental melts, endorsing a view¹⁷ that intermediate melts often result from magma mixing. This view is corroborated by the fact that most suites of intermediate rocks describe linear chemical trends, rather than the curved trends generated experimentally. It is striking that the mismatch between the experimental liquid lines of descent and intermediate bulk rocks exactly corresponds to the compositional gap recorded by the melt inclusions, leaving little doubt that the vast majority of intermediate arc magmas do not represent melts produced by differentiation of mafic magmas.

The melt inclusion record (Fig. 1c, d) shows that andesitic volcanoes contain almost exclusively dacitic to rhyolitic liquids. The majority of intermediate rocks lack mafic melt inclusions, which raises questions as to what drives their bulk composition back to andesite. The most commonly invoked process is magma mixing, for which there is abundant petrological evidence in andesitic rocks¹⁷. However, the paucity of andesitic melt inclusions suggests that mixing itself does not produce large volumes of hybrid andesitic melts. At Volcán de Colima (Mexico), for example, there are strong indications that the linear andesitic bulk-rock trends are controlled by incorporation of gabbroic fragments into silicic melts, as represented by melt inclusions¹⁸. The gabbroic fragments are themselves a product of crystallization of ancestral mafic magmas beneath the volcano. The presence of disequilibrium mafic crystals and plutonic nodules is a well known feature of intermediate arc magmas¹⁹, suggesting that mechanical incorporation of mafic rock fragments into silicic melts may be a widespread process. Incorporation could result either from assimilation of mafic plutonic roots by ascending silicic melts¹⁹ or incomplete separation of restitic fragments and residual silicic melts during extraction from their source regions²⁰. Given the abundance of plutonic nodules at many volcanic arcs, it is axiomatic that most ascending magmas encounter and interact with ancestral crystalline residues before eruption.

Considering that melt inclusions record upper crustal pressures (≤ 550 MPa), the compositional gap must be produced during an early stage of magmatic evolution, in the mid to deep crust, implying that melt compositional diversity is acquired at depth, as suggested in refs 16 and 21. In view of the good agreement between the experimental liquid lines of descent of hydrous basalts and silicic melt inclusions (Fig. 3), the compositional gap is probably related to the formation and/or extraction of silicic melts during crystallization of basalts and/or melting of mafic source rocks in the mid to deep crust.

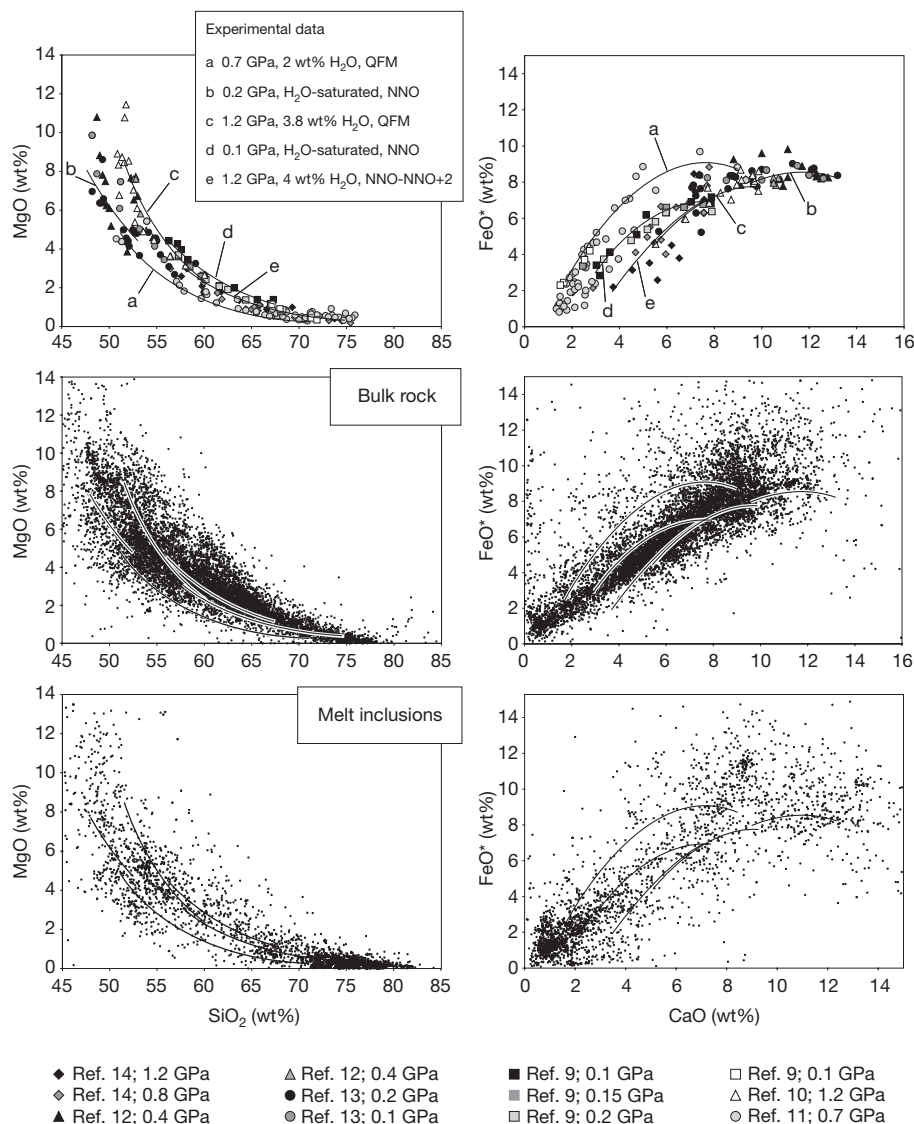


Figure 3 | Comparison of chemical variation in melt inclusions, volcanic bulk rocks and experimental melts. The represented liquid lines of descent (curves a–e in top panels, shown in panels below) are based on experimental studies of hydrous basalts over a range of crustal pressures (0.1–1.2 GPa). Bulk rock data (symbols) are from a GEOROC compilation for North, Central and South America. Melt inclusions as in Fig. 1. The mismatch

Melts produced experimentally by crystallization of hydrous basalts span a continuous range of composition without any obvious gap (Fig. 3). However, over a wide range of crustal pressures, the slope of the liquidus in temperature–composition space shallows markedly between 58–60 and 64–68 wt% SiO₂, which correlates well with the gap recorded by the melt inclusions (Fig. 4). As originally proposed²², large compositional changes over small temperature intervals significantly reduce the likelihood of extracting compositions within these intervals compared to compositions corresponding to steeper portions of the liquidus at higher and lower SiO₂. We propose that the andesitic gap shown by the melt inclusion record is a consequence of the phase relations of hydrous basalts in the mid to lower crust. Both evolved and basaltic compositions ascend and crystallize in the shallow crust: the production of the vast majority of andesitic rocks is simply a consequence of the blending of these components in upper crustal reservoirs.

Recognizing the existence of a gap in melt composition and the mixed nature of most erupted andesites solves several puzzling aspects of arc magmatism. A review⁸ of Quaternary magmatism in the continental Cascades arc (USA) highlights the fact that the vast

observed between the experimental liquid lines of descent and many intermediate bulk rocks, which are displaced to higher MgO relative to experimental liquids, coincides with the compositional gap recorded by the melt inclusions (see Fig. 1), strongly suggesting rarity of true intermediate melts in arc settings.

majority of intermediate magmas (57–68 wt% SiO₂) were erupted from 22 major stratovolcanoes, whereas more than 2,000 single vents dispersed throughout the arc produced mafic magmas (<60 wt% SiO₂). That the predominant mafic cones and shield volcanoes do not grade into high silica andesites is difficult to explain if silicic melts were produced by low-pressure basalt crystallization, but is consistent with a melt inclusion record showing that, in most cases, vapour-saturated crystallization of basaltic melt is limited to producing rocks with less than 60 wt% SiO₂. The review in ref. 8 also shows that, in contrast to intermediate magmas, whose occurrence is restricted to major stratovolcanoes, silicic magmas were erupted from independent peripheral vents and dome fields. With the exception of the 22 major stratovolcanoes, the Cascades therefore shows evidence for a clear bimodality. Recent studies of oceanic arc bulk rocks have also shown distinct bimodality, with minima at high silica andesite^{23,24}. This concurs with the melt inclusion and experimental records, indicating that intermediate liquids are not typically primary products of the subduction factory.

The link between plutonic and volcanic rocks is the key to understanding the formation and evolution of magmatic systems. Recent

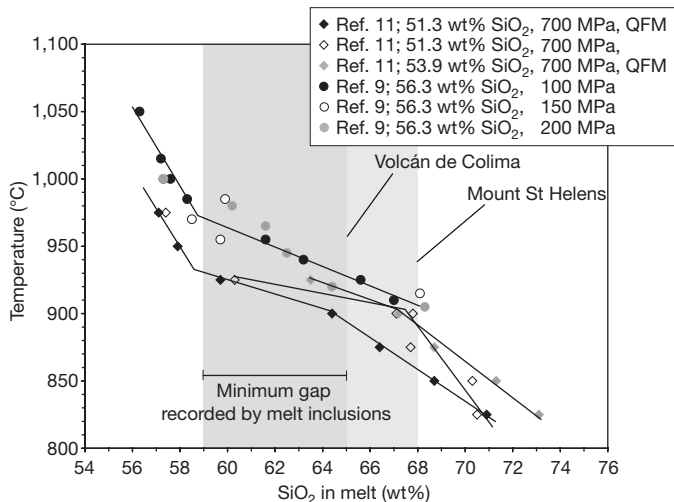


Figure 4 | Experimental liquid lines of descent in temperature-composition space at a range of crustal pressures (100–700 MPa). Grey areas represent the compositional gap observed in the melt inclusion record. Volcán de Colima marks the minimum extension of this gap, whereas Mount St Helens represents the common width of the gap at arc volcanoes (see Fig. 1b). The SiO₂ contents in the key are starting melt compositions used in experiments.

studies have presented several lines of evidence that plutonic bulk rock compositions are in large part representative of solidified melts rather than crystal cumulates^{25,26}. Good correlations between melt inclusions, experimental liquid lines of descent and plutonic rock compositions support this hypothesis for mafic and silicic compositions (Fig. 5). However, intermediate plutonic rocks, like their volcanic counterparts, appear to be products of mingling and mixing. Plutons commonly show clear field evidence for bimodality of incoming melts and subsequent mechanical and/or chemical interaction between these components²⁷, supporting an upper crustal mixing/mingling origin for intermediate plutonic rocks. The most striking discrepancies between the volcanic and plutonic records are

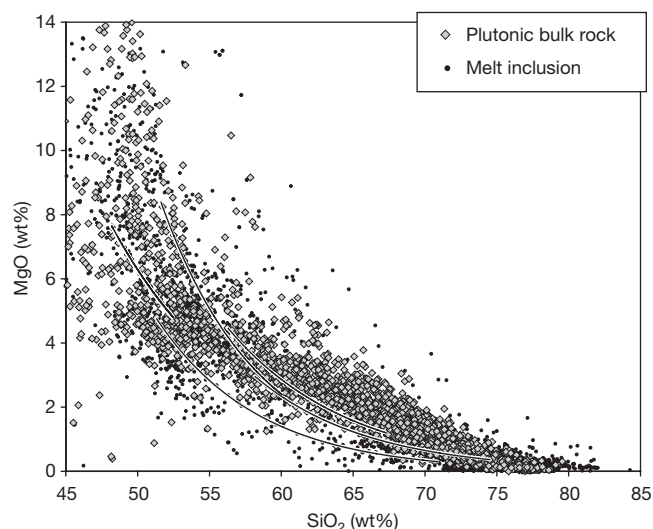


Figure 5 | Chemical variation in melt inclusions, plutonic bulk rocks and experimental melts. Diamonds are plutonic bulk rock compositions (data set as in Fig. 1a), small dots represent the GEOROC melt inclusion compilation for arc volcanoes, curves are experimental liquid lines of descent as in Fig. 3. Good correlation between bulk rock and melt inclusion compositions suggests that mafic and silicic plutonic rocks essentially represent melts. As with the volcanic rocks (Fig. 3), intermediate plutonic rocks show a mismatch with experimental melts, which corresponds to the gap recorded by melt inclusions, demonstrating that true intermediate melts are rare in plutons.

the paucity of andesite compositions and the large amount of silicic magmas in plutons (Figs 1a, 5). These discrepancies may be reconciled if we consider the constraints provided by the melt inclusion record, namely that intermediate magmas form by mingling/mixing process within bimodal magmatic systems and large volumes of dacitic to rhyolitic melts occur beneath andesitic volcanoes. In addition, plutons often lack the types of internal compositional zoning that would be consistent with *in situ* differentiation, suggesting, once again, that upper crustal magmatic systems are the site of extensive crystallization and blending between bimodal melts ascending from the lower crust rather than fractionation-driven differentiation.

METHODS SUMMARY

Melt inclusion data for arc volcanoes were extracted from the GEOROC melt inclusion global compilation (precompiled file available at <http://georoc.mpch-mainz.gwdg.de/georoc/>). Bulk rock compositions shown in figures were also obtained from GEOROC precompiled files. Cretaceous plutonic bulk rock compositions for the Sierra Nevada batholith were extracted from the NAVDAT compilation (<http://www.navdat.org/>). All data presented are calculated on an H₂O-free basis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 19 May; accepted 15 September 2009.

- Rudnick, R. L. Making continental crust. *Nature* **378**, 571–578 (1995).
- Gill, J. B. *Orogenic Andesites and Plate Tectonics* (Springer, 1981).
- Naumov, B. V., Kovalenko, V. I., Babansky, A. D. & Tolstykh, M. L. Genesis of andesites: evidence from studies of melt inclusions in minerals. *Petrology* **5**, 586–596 (1997).
- Blundy, J. & Cashman, K. Rapid decompression-driven crystallization recorded by melt inclusions from Mount St. Helens volcano. *Geology* **33**, 793–796 (2005).
- Wallace, P. J. Volatiles in subduction zone magmas: concentrations and fluxes based on melt inclusion and volcanic gas data. *J. Volcanol. Geotherm. Res.* **140**, 217–240 (2005).
- Martin, V. M., Holness, M. B. & Pyle, D. M. Textural analysis of magmatic enclaves from the Kameni Islands, Santorini, Greece. *J. Volcanol. Geotherm. Res.* **154**, 89–102 (2006).
- Naranjo, J. A., Sparks, R. S. J., Stasiuk, M. V., Moreno, H. & Ablay, G. J. Morphological, structural and textural variations in the 1988–1990 andesite lava of Lonquimay Volcano, Chile. *Geol. Mag.* **129**, 657–678 (1992).
- Hildreth, W. Quaternary magmatism in the Cascades — geologic perspectives. *Prof. Pap. US Geol. Surv.* **1744** (2007).
- Grove, T. L., Donnelly-Nolan, J. M. & Housh, T. Magmatic processes that generated the rhyolite of Glass Mountain, Medicine Lake volcano, N California. *Contrib. Mineral. Petrol.* **127**, 205–223 (1997).
- Müntener, O., Kelemen, P. B. & Grove, T. L. The role of H₂O during crystallization of primitive arc magmas under uppermost mantle conditions and genesis of igneous pyroxenites: an experimental study. *Contrib. Mineral. Petrol.* **141**, 643–658 (2001).
- Sisson, T. W., Ratajeski, K., Hanks, W. B. & Glazner, A. F. Voluminous granitic magmas from common basaltic sources. *Contrib. Mineral. Petrol.* **148**, 635–661 (2005).
- Pichavant, M. & Macdonald, R. Crystallization of primitive basaltic magmas at crustal pressures and genesis of the calc-alkaline igneous suite: experimental evidence from St Vincent, Lesser Antilles arc. *Contrib. Mineral. Petrol.* **154**, 535–558 (2007).
- Sisson, T. W. & Grove, T. L. Experimental investigations of the role of H₂O in calc-alkaline differentiation and subduction zone magmatism. *Contrib. Mineral. Petrol.* **113**, 143–166 (1993).
- Alonso-Perez, R., Müntener, O. & Ulmer, P. Igneous garnet and amphibole fractionation in the roots of island arcs: experimental constraints on andesitic liquids. *Contrib. Mineral. Petrol.* **157**, 541–558 (2009).
- Depaolo, D. J., Perry, F. V. & Baldrige, W. S. Crustal versus mantle sources of granitic magmas — a 2-parameter model based on Nd isotopic studies. *Trans. R. Soc. Edinb.* **83**, 439–446 (1992).
- Hildreth, W. & Moorbath, S. Crustal contributions to arc magmatism in the Andes of central Chile. *Contrib. Mineral. Petrol.* **98**, 455–489 (1988).
- Anderson, A. T. Magma mixing — petrological process and volcanological tool. *J. Volcanol. Geotherm. Res.* **1**, 3–33 (1976).
- Reubi, O. & Blundy, J. Assimilation of plutonic roots, formation of high-K exotic melts and genesis of andesitic magmas at Volcán de Colima, Mexico. *J. Petrol.* **49**, 2221–2243 (2008).
- Dungan, M. A. & Davidson, J. Partial assimilative recycling of the mafic plutonic roots of arc volcanoes: an example from the Chilean Andes. *Geology* **32**, 773–776 (2004).
- Chappell, B. W. & White, A. J. R. Two contrasting granite types: 25 years later. *Aust. J. Earth Sci.* **48**, 489–499 (2001).

21. Annen, C., Blundy, J. D. & Sparks, R. S. J. The genesis of intermediate and silicic magmas in deep crustal hot zones. *J. Petrol.* **47**, 505–539 (2006).
22. Grove, T. L. & Donnelly-Nolan, J. M. The evolution of young silicic lavas at Medicine Lake Volcano, California — implications for the origin of compositional gaps in calc-alkaline series lavas. *Contrib. Mineral. Petrol.* **92**, 281–302 (1986).
23. Arculus, R. J. in *State of the Arc Conference Extended Abstracts and Programme* (eds Dungan, M. A., Grunder, A., Hickey-Vargas, R., Moreno Roa, H. & Muñoz, J.) 1–4 (IAVCEI, 2007).
24. Wright, I. C. & Gamble, J. A. Southern Kermadec submarine caldera arc volcanoes (SW Pacific): caldera formation by effusive and pyroclastic eruption. *Mar. Geol.* **161**, 207–227 (1999).
25. Glazner, A. F., Bartley, J. M. & Coleman, D. S. in *State of the Arc Conference Extended Abstracts and Programme* (eds Dungan, M. A., Grunder, A., Hickey-Vargas, R., Moreno Roa, H. & Muñoz, J.) 79–80 (IAVCEI, 2007).
26. Ulmer, P. Differentiation of mantle-derived calc-alkaline magmas at mid to lower crustal levels: experimental and petrologic constraints. *Period. Mineral.* **76**, 309–325 (2007).
27. Blundy, J. D. & Sparks, R. S. J. Petrogenesis of mafic inclusions in granitoids of the Adamello Massif, Italy. *J. Petrol.* **33**, 1039–1104 (1992).
28. Blundy, J., Cashman, K. & Humphreys, M. Magma heating by decompression-driven crystallization beneath andesite volcanoes. *Nature* **443**, 76–80 (2006).
29. Humphreys, M. C. S., Blundy, J. D. & Sparks, R. S. J. Shallow-level decompression crystallisation and deep magma supply at Shiveluch Volcano. *Contrib. Mineral. Petrol.* **155**, 45–61 (2008).
30. Johnson, E. R., Wallace, P. J., Cashman, K. V., Granados, H. D. & Kent, A. J. R. Magmatic volatile contents and degassing-induced crystallization at Volcán Jorullo, Mexico: implications for melt evolution and the plumbing systems of monogenetic volcanoes. *Earth Planet. Sci. Lett.* **269**, 477–486 (2008).
31. Roggensack, K. Unraveling the 1974 eruption of Fuego volcano (Guatemala) with small crystals and their young melt inclusions. *Geology* **29**, 911–914 (2001).

Acknowledgements This work was supported by a Marie Curie Fellowship (O.R.) and an NERC Senior Research Fellowship (J.B.). We thank S. Sparks, K. Kelley and K. Roggensack for comments on an early draft of this manuscript, and P. Wallace and M. Pichavant for critical reviews. We are grateful to V. C. Smith for unpublished data and discussions.

Author Contributions O.R. and J.B. developed the discussion. O.R. took the lead in writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to O.R. (olivier.reubi@erdw.ethz.ch).

METHODS

Modification of melt inclusions after trapping, either through crystallization of the host phase onto inclusion walls or by growth of daughter crystals, is an enduring concern in the interpretation of melt inclusion data (see, for example, ref. 32). Typically, the compositions of mafic melt inclusions in olivine are corrected for post-entrapment crystallization of the host phase and diffusive loss of Fe (refs 33, 34) whereas silicic melt inclusions are uncorrected, potentially producing a bias in the data set presented. However, the extent of the correction applied to mafic melt inclusions is small (generally <2 wt% SiO₂) and cannot generate a compositional gap. Detailed studies of silicic melt inclusions in andesites from Volcán de Colima (Mexico), Mount St Helens (USA) and Shiveluch (Russia)^{4,18,29} have shown that melt inclusions show the same range of composition irrespective of their host mineral. Post-entrapment re-equilibration of melt inclusions would result in compositional evolution controlled by crystallization of the host or diffusion through the host. In both cases, the melt inclusion will have a chemistry that is controlled by the crystal–melt partition coefficients and/or relative diffusivity in the host crystal³⁵. Compositions modified by post-entrapment processes are, of necessity, specific to their host mineral. Thus melt inclusions in plagioclase, pyroxenes, amphibole and quartz should form distinct populations; this is demonstrably not the case, as can be seen from Figs 1 and 3. For example, if post-entrapment crystallization played a significant role, then we would see a difference in SiO₂ contents of melts trapped in amphibole (a low-SiO₂ phase) versus orthopyroxene (a high-SiO₂ phase). This is not observed, and the shift towards more silicic compositions shown by the melt inclusions compared to host rock compositions is observed for plagioclase, orthopyroxene and amphibole (Fig. 1) and therefore cannot result from specific entrapment or preservation mechanisms.

An additional concern is the ability of melt inclusions to record the evolution of the melt throughout the course of crystallization. Studies of andesitic volcanoes have shown that bulk rocks and melt inclusions form oblique chemical trends that intersect at, or close to, the composition of the least evolved melt inclusions^{4,18,29}. This clearly demonstrates that at these volcanoes, at least, the melt inclusions effectively record the composition of the melt from close to the onset of low pressure crystallization (that is, to within 1–2 wt% SiO₂). This may sound surprising, considering that crystals need to be growing from melt before they can trap melt inclusions. However the presence of older inherited crystals (antecrysts) is a well known feature of intermediate rocks, and it is likely that at least some melt inclusions become trapped within antecrysts incorporated in ascending melts. Indeed, resorption of antecrysts in superheated hydrous melts ascending through the crust²¹ and subsequent low-pressure overgrowth rims formed as the melts reach their H₂O-saturated liquidus is an efficient melt-inclusion-forming mechanism¹⁸.

It is often assumed that melt inclusions out of chemical equilibrium with their host mineral are unrepresentative of the surrounding melt and form either from

grain boundary processes or post entrapment re-equilibration. However, resorption is one of several possible mechanism of melt inclusion formation³² and the existence of melt inclusions out of equilibrium with the composition of the host adjacent to them is to be expected, as discussed above. Detailed compositional mapping of the host crystal zoning pattern and consideration of three-dimensional effects are often essential to establish the melt inclusion entrapment mechanism. Silicic melt inclusions in resorbed An-rich plagioclase^{4,36} are one such example. Careful evaluation of the melt inclusion chemistry, comparison between melt inclusions in different host minerals and with host rock and groundmass compositions is essential to establish representativeness. Disequilibrium between melt inclusions and their host crystal does not itself signify that the composition of the melt inclusion is not representative of the melt, it is an indication of the mechanism of formation and entrapment.

Melt inclusions with compositions clearly distinct from host rock and groundmass glass compositions have been reported in arc magmas^{18,37}. These ‘exotic’ melt inclusions are interpreted as reflecting localized, grain-scale dissolution processes in crystal cumulates and cannot be taken as representative of the composition of the crystallizing melt. However, because the compositions of the vast majority of melt inclusions are consistent with bulk rock, experimental melt and/or groundmass glass compositions in both mafic and silicic magmas, it appears that exotic melt inclusions are relatively rare and there is no doubt that most melt inclusions effectively record the evolution of the melts feeding arc volcanoes.

We conclude that neither post-entrapment re-equilibration of melt inclusions, crystallization of the host phase nor bias in the melt inclusion record can explain the observed compositional gap and the occurrence of a silicic peak. Consequently these features represent primary characteristics of arc magmatism.

32. Roedder, E. *Fluid Inclusions* (Mineralogical Society of America, 1984).
33. Danyushevsky, L. V., Della-Pasqua, F. N. & Sokolov, S. Re-equilibration of melt inclusions trapped by magnesian olivine phenocrysts from subduction-related magmas: petrological implications. *Contrib. Mineral. Petrol.* **138**, 68–83 (2000).
34. Danyushevsky, L. V., McNeill, A. W. & Sobolev, A. V. Experimental and petrological studies of melt inclusions in phenocrysts from mantle-derived magmas: an overview of techniques, advantages and complications. *Chem. Geol.* **183**, 5–24 (2002).
35. Cottrell, E., Spiegelman, M. & Langmuir, C. H. Consequences of diffusive reequilibration for the interpretation of melt inclusions. *Geochem. Geophys. Geosyst.* **3**, 1–26 (2002).
36. Humphreys, M. C. S., Blundy, J. D. & Sparks, R. S. J. Magma evolution and open-system processes at Shiveluch Volcano: insights from phenocryst zoning. *J. Petrol.* **47**, 2303–2334 (2006).
37. Danyushevsky, L. V., Leslie, R. A. J., Crawford, A. J. & Durance, P. Melt inclusions in primitive olivine phenocrysts: the role of localized reaction processes in the origin of anomalous compositions. *J. Petrol.* **45**, 2531–2553 (2004).

LETTERS

Visual but not trigeminal mediation of magnetic compass information in a migratory bird

Manuela Zapka¹, Dominik Heyers¹, Christine M. Hein¹, Svenja Engels¹, Nils-Lasse Schneider¹, Jörg Hans¹, Simon Weiler¹, David Dreyer¹, Dmitry Kishkinev¹, J. Martin Wild² & Henrik Mouritsen¹

Magnetic compass information has a key role in bird orientation^{1–3}, but the physiological mechanisms enabling birds to sense the Earth's magnetic field remain one of the unresolved mysteries in biology^{2,4}. Two biophysical mechanisms have become established as the most promising magnetodetection candidates. The iron-mineral-based hypothesis suggests that magnetic information is detected by magnetoreceptors in the upper beak and transmitted through the ophthalmic branch of the trigeminal nerve to the brain^{5–10}. The light-dependent hypothesis suggests that magnetic field direction is sensed by radical pair-forming photopigments in the eyes^{11–15} and that this visual signal is processed in cluster N, a specialized, night-time active, light-processing forebrain region^{16–19}. Here we report that European robins with bilateral lesions of cluster N are unable to show oriented magnetic-compass-guided behaviour but are able to perform sun compass and star compass orientation behaviour. In contrast, bilateral section of the ophthalmic branch of the trigeminal nerve in European robins did not influence the birds' ability to use their magnetic compass for orientation. These data show that cluster N is required for magnetic compass orientation in this species and indicate that it may be specifically involved in processing of magnetic compass information. Furthermore, the data strongly suggest that a vision-mediated mechanism underlies the magnetic compass in this migratory songbird, and that the putative iron-mineral-based receptors in the upper beak connected to the brain by the trigeminal nerve^{6–8} are neither necessary nor sufficient for magnetic compass orientation in European robins.

Thirty-six European robins (*Erithacus rubecula*) were caught within 200 m of the testing site and their spontaneous migratory orientation was tested in modified Emlen funnels^{20,21} inside wooden huts in the natural magnetic field (NMF) and in a magnetic field with geomagnetic north turned horizontally 120° anticlockwise (CMF). After these control tests confirmed well-oriented magnetic compass behaviour, one of the following surgeries was performed: bilateral section ($N = 7$; for details, see Methods) of the ophthalmic branch of the trigeminal nerve (V_1); trigeminal sham section ($N = 6$; the same treatment except that V_1 was not sectioned); chemical lesion of cluster N ($N = 13$; bilateral, focal injections of ibotenic acid were made into cluster N); or sham lesion of cluster N ($N = 10$; the same treatment but without injection of ibotenic acid). The surgeries were performed by two of us (J.M.W. and D.H.) without the others knowing which bird underwent which surgery. After the surgery and a recovery period of at least one week, the birds were retested under the same magnetic conditions as in the control tests (NMF and CMF) during spring migration. Subsamples of the cluster-N-lesioned or sham-lesioned birds were also tested in a natural magnetic field with an inverted vertical component (IMF). All orientation results were evaluated independently by two or three individuals who did not

know which kind of surgery (real or sham) or which magnetic condition the birds had experienced.

The orientation results showed that in European robins the ophthalmic branch of the trigeminal nerve is not necessary (Fig. 1) for magnetic compass orientation, whereas cluster N is necessary (Fig. 2a–f). The sham-sectioned birds oriented north in the geomagnetic field ($\alpha = 10^\circ \pm 20^\circ$ (mean vector orientation angle; 95% confidence interval), $r = 0.95$ (mean vector length), $N = 6$, $P < 0.002$; Fig. 1a). When geomagnetic north was turned to 240°, the same birds oriented west-southwest ($\alpha = 245^\circ \pm 26^\circ$, $r = 0.90$, $N = 6$, $P < 0.005$; Fig. 1b). The trigeminal-sectioned birds also clearly oriented north in the geomagnetic field ($\alpha = 354^\circ \pm 20^\circ$, $r = 0.93$, $N = 7$, $P < 0.001$; Fig. 1d) and west-southwest when geomagnetic north was turned to 240° ($\alpha = 264^\circ \pm 34^\circ$, $r = 0.77$, $N = 7$, $P < 0.01$; Fig. 1e). The 95% confidence intervals in Fig. 1a, b and Fig. 1d, e do not overlap, so we conclude that both groups significantly changed their orientation in response to the turned magnetic field. Furthermore, the mean orientation of both groups under the CMF condition was not significantly different from

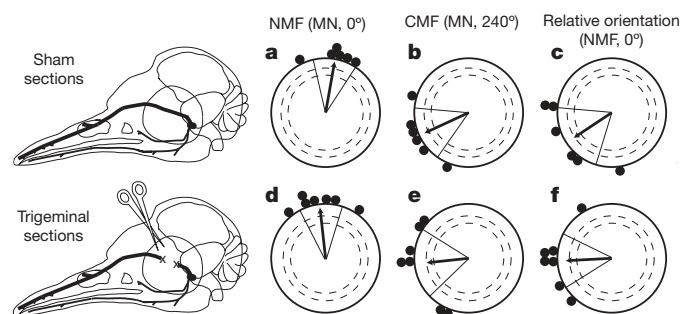


Figure 1 | Bilateral sectioning of the ophthalmic branch of the trigeminal nerve does not affect magnetic compass orientation in European robins. The drawings show the approximate locations of the three branches of the trigeminal nerve. The ophthalmic branch (V_1) is shown in bold. The crosses indicate the approximate locations at which the nerve was sectioned and a piece removed. **a–c**, Magnetic orientation of six sham-sectioned birds (MN, magnetic north). **d–f**, Magnetic orientation of seven trigeminal-sectioned birds. Each filled circle at the periphery indicates the mean orientation of an individual bird based on nine tests under the given magnetic condition. **c** and **f** compare the orientation of each bird in the turned magnetic field (CMF) with the same bird's orientation in the natural magnetic field (NMF, standardized to 0°). Arrows indicate the group mean vectors. The longer is the group mean vector, the more consistent are the orientation choices between individuals. Inner and outer dashed circles indicate the radii of the group mean vectors needed for directional significance according to the Rayleigh test (inner, $P < 0.05$; outer, $P < 0.01$). Radial lines flanking the group mean vector mark the 95% confidence interval for the group mean direction.

¹AG Neurosensorik/Animal Navigation, IBU, University of Oldenburg, D-26111 Oldenburg, Germany. ²Department of Anatomy, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand.

that expected to result from a magnetic field turn of -120° , but was significantly different from the same birds' orientation under the NMF condition (the 95% confidence intervals of Fig. 1c, f include 240° but do not include 0°).

The sham-lesioned birds oriented north-northeast ($\alpha = 25^\circ \pm 21^\circ$, $r = 0.90$, $N = 10$, $P < 0.001$; Fig. 2a) in the geomagnetic field. When geomagnetic north was turned to 240° , the same birds oriented southwest ($\alpha = 219^\circ \pm 26^\circ$, $r = 0.83$, $N = 10$, $P < 0.001$; Fig. 2b), and when the vertical component of the geomagnetic field was inverted, the same birds oriented south-southeast ($\alpha = 154^\circ \pm 37^\circ$, $r = 0.69$, $N = 10$, $P < 0.01$; Fig. 2c). None of these 95% confidence intervals overlap, so we conclude that the birds oriented significantly differently under the three magnetic conditions. Judging from the 95% confidence intervals, the orientations of the CMF and IMF groups were not significantly different from those expected to result from magnetic field turns of -120° and 180° .

In contrast to the well-oriented, sham-lesioned birds, the birds with a chemical lesion of cluster N oriented randomly under all of the three magnetic field conditions (NMF: $\alpha = 355^\circ$, $r = 0.33$, $N = 13$, $P = 0.25$; CMF: $\alpha = 318^\circ$, $r = 0.40$, $N = 13$, $P = 0.13$; IMF: $\alpha = 120^\circ$, $r = 0.07$, $N = 7$, $P > 0.90$; Fig. 2d–f). Furthermore, the consistency of the birds' directional choices between tests was significantly poorer in the cluster-N-lesioned birds than in the sham-lesioned birds (comparing the r values for the individual mean directions in NMF: t -test, $N_{\text{lesion}} = 13$, $N_{\text{sham}} = 10$, $t = 3.160$, $P < 0.01$). Consequently, birds with lesions of cluster N cannot perform magnetic compass orientation in an orientation cage. The fact that the cluster-N-lesioned birds, which possessed intact trigeminal nerves, did not orient indicates that information transmitted through the ophthalmic branch of the trigeminal nerve is not sufficient for magnetic compass orientation in European robins.

After completion of the experiments, all birds that had undergone real (non-sham) surgery were killed for anatomical/histological analysis. In all trigeminal-sectioned specimens, the nerves were found not to have rejoined. Brains from cluster-N-lesioned birds were sectioned and stained for anti-human neuronal protein²²

(anti-HuC (also known as anti-ELAVL3) or anti-HuD (anti-ELAVL4), Molecular Probes), which enabled us to determine the extent of the lesions. Most birds had well-placed lesions covering at least 66% of cluster N on both sides of the brain (mean \pm s.d., $78 \pm 9\%$). In three of the 13 lesioned birds, less than 50% of cluster N was lesioned (no. 14: 63% of left side, 33% of right side; no. 16: 17% of left side, 24% of right side; no. 18: 35% of left side, 29% of right side). We note that these three birds oriented well under all three magnetic field conditions ($N = 3$; NMF: $\alpha = 27^\circ$, $r = 0.82$; CMF: $\alpha = 249^\circ$, $r = 0.97$; IMF: $\alpha = 184^\circ$, $r = 0.81$; mean vectors lie within the 99% confidence intervals of the mean orientation of the sham-lesioned birds under all magnetic field conditions; Fig. 2d–f, open circles).

Our results, and the fact that cluster N is part of the visual system^{16–19}, appear to strongly support the hypothesis that magnetic compass input is processed in the visual system of night-migratory passerines. However, we consider two alternative explanations. The first is reduced general night-vision capability. Considering that cluster N is part of the visual Wulst¹⁸ and that it is known to process night-time light-dependent information^{16,17}, we tested whether a reduction of general night-vision capability could explain the difference in orientation performance between cluster-N-lesioned and sham-lesioned birds. First, we noticed that the cluster-N-lesioned birds also showed a high level of migratory restlessness in the funnels (346 ± 184 (mean \pm s.d.) scratches per hour and per active test). In contrast, when we tested European robins in complete darkness, that is, when they were unable to see, they showed very little migratory restlessness (on average < 20 scratches per hour). Second, Wulst lesions in pigeons (*Columba livia*) have been shown to affect the threshold for detecting the intensity of a dim point of light²³. Therefore, we performed operant conditioning tests in which European robins, one group with cluster N lesions and one group without, were trained to detect a dim point of light. Both groups could perform this visual discrimination task at light intensities 400 times dimmer than the light level under which the magnetic compass orientation tests were

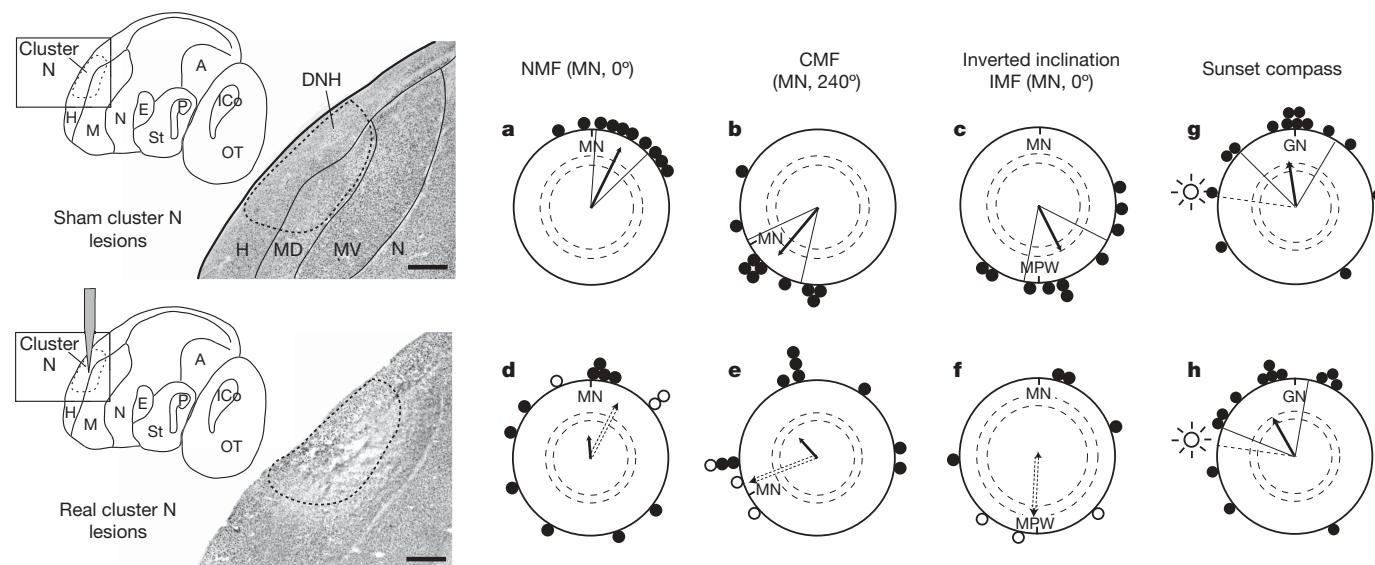


Figure 2 | Bilateral lesions of cluster N disrupt magnetic compass orientation in European robins. The photos show examples of brain sections from a sham-lesioned (top) and an actually lesioned (bottom) European robin, each sagittally cut through the centre of cluster N and stained with anti-HuC/HuD (a neuronal marker). The tissue where cluster N should have been in the lesioned bird is destroyed (compare with top photo). The drawings indicate where in the brain the photos were taken: A, arcopallium; E, entopallium; H, hyperpallium; DNH, dorsal nucleus of the hyperpallium; ICo, intercollicular complex; M, mesopallium (MD, mesopallium dorsale; MV, mesopallium ventrale); N, nidopallium; OT, optic tectum; P, pallidum; St, striatum. Rostral, left; caudal, right. Scale bar, 500 μ m. **a–c**, Well-

oriented, springtime, magnetic compass behaviour of the ten birds that received sham cluster N lesions (MPW, direction of magnetic pole for an inclination compass). **d–f**, Non-oriented magnetic compass behaviour of the 13 cluster-N-lesioned birds. The data are means of 14 tests of each individual in the NMF, 13 tests in the CMF and 8 tests in the IMF. The open circles in **d–f** show the mean orientations of the three birds in which only 20%, 32% or 47%, respectively, of cluster N was lesioned. The dashed arrows show the corresponding mean vectors. **g, h**, Sunset orientation (GN, geographical north) of the sham-lesioned (**g**) and lesioned (**h**) birds (14 tests per bird). The dashed radial line indicates the average sunset direction during the tests. Other details of the circular diagrams are as in Fig. 1.

performed (the visual detection limit was $<0.01 \text{ mW m}^{-2}$ for both groups; Supplementary Fig. 1). We therefore conclude that a lesion-induced reduction in visual capability is unlikely to have caused the observed differences in orientation performance.

The second alternative explanation is that cluster N might not specifically be involved in the circuit processing magnetic compass information. Finding directions in a magnetic orientation experiment is a complex task involving coordination of information from multiple neural pathways and probably input from memory. Lesions of cluster N could have affected neural processes required for motivation to migrate or for solving orientation tasks in general, thus leading to disorientation even though magnetic sensing itself remains intact. To test the specificity of cluster N in magnetic compass orientation, we tested 13 cluster-N-lesioned European robins and 14 sham-lesioned European robins for their orientation abilities under two conditions: outdoors during sunset and indoors under a stationary planetarium sky simulating the local starry sky (Oldenburg, Germany; Figs 2 and 3).

During natural sunset, both the sham-lesioned and cluster-N-lesioned birds oriented significantly towards the north-northwest (sham-lesioned group: $\alpha = 353^\circ \pm 38^\circ$, $r = 0.57$, $N = 14$, $P < 0.01$; cluster-N-lesioned group: $\alpha = 331^\circ \pm 39^\circ$, $r = 0.57$, $N = 13$, $P = 0.01$; Fig. 2g, h). The mean orientation of both groups indicates a compromise direction between phototactic tendencies towards the setting sun and the birds' north-northeast migratory direction. This reaction is typical of outdoor sun compass orientation tests on migratory birds during sunset²⁴. The point is that the birds' orientation was significantly more northerly than the sunset direction, meaning that pure phototactic orientation can be excluded (95% confidence intervals for the mean direction do not overlap with the sunset point, which on average was at 278° during our experiments).

In the planetarium, we simulated celestial north to be located at magnetic east so that we could determine whether orientation behaviour was guided by a star compass or a magnetic compass. To encourage the birds to use their star compass, we added more magnetic disturbance to the already significantly disturbed geomagnetic field inside the planetarium (Methods). The cluster-N-lesioned birds oriented significantly towards star north-northeast ($\alpha = 27^\circ \pm 44^\circ$, $r = 0.55$, $N = 12$, $P = 0.02$; Fig. 3a) and the mean orientation was almost identical to the direction chosen by the sham-lesioned birds using their magnetic compass ($\alpha = 25^\circ \pm 21^\circ$; Fig. 2a). These results show that birds with bilateral lesions of cluster N can use their star compass and their sunset compass, but cannot use their magnetic compass to perform appropriately directed migratory restlessness behaviour (Fig. 3). Therefore, a generally reduced motivation or inability to migrate cannot explain the disorientation of birds with cluster N lesions.

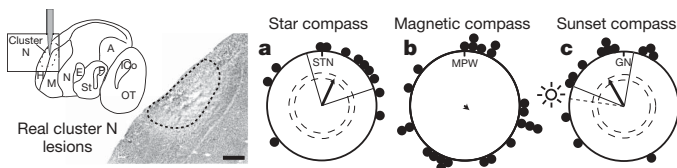


Figure 3 | Birds with cluster N lesions can use their star and sun compasses, but not their magnetic compass. Photo from a lesioned bird as in Fig. 2. **a**, Birds with cluster N lesions tested in a planetarium simulating the local starry sky (ten tests per bird; STN, star north) oriented in the typical north-northeast spring migratory direction ($\alpha = 27^\circ \pm 44^\circ$, $r = 0.55$, $N = 12$, $P = 0.02$). **b**, Birds with cluster N lesions could not orient ($\alpha = 132^\circ$, $r = 0.12$, $P > 0.70$) using their magnetic compass. Shown are the combined data of Fig. 2d–f, depicted relative to the magnetic direction towards the pole (MPW) as European robins use an inclination compass (data from the three poorly lesioned birds are not included). **c**, Birds with cluster N lesions could also orient during sunset ($\alpha = 331^\circ \pm 39^\circ$, $r = 0.57$, $N = 13$, $P = 0.01$), presumably using their sun compass.

Because the ophthalmic branch of the trigeminal nerve is the only nerve branch that innervates the candidate ferromagnetic, magneto-sensory structures in the upper beak^{7–9}, our results show that these putative magnetoreceptors are neither necessary nor sufficient for magnetic compass orientation, and can therefore be excluded as the sole magnetic compass sensor in European robins. This conclusion is in line with results from other studies including, for instance, those on bobolinks, in which anaesthetic blockade of the trigeminal nerve also failed to affect compass orientation²⁵. However, our findings do not rule out the possibility that these or other putative magnetoreceptors in other regions of the body can sense geomagnetic information^{26,27}. In fact, in pigeons, the putative magnetosensors in the upper beak have been strongly implicated in the sensing of non-compass aspects of the geomagnetic field¹⁰ (but see also refs 28, 29).

The results of the present study, together with those which show that cluster N is the most active forebrain region during magnetic compass orientation behaviour^{16,17,19} and is a specialized part of the visual system¹⁸ requiring light perceived through the eyes for its neuronal activation^{16,17}, specifically suggest that cluster N of European robins is an essential part of a circuit processing light-dependent magnetic compass information for night-time orientation. The exact role of cluster N within this circuit has not been determined, but the present results raise the distinct possibility that this part of the visual system enables birds to 'see' magnetic compass information.

METHODS SUMMARY

All magnetic field conditions were produced using double-wrapped, three-dimensional Merritt four-coil systems³⁰ with average coil diameters of 2 m. All experiments were performed within the central space of the coils, where the heterogeneities were $<1\%$ of the applied field. Current flowed through the coils in all magnetic conditions.

The operations were performed under general anaesthesia. To lesion cluster N, 50 nl 1% ibotenic acid in 0.9% NaCl were injected with a microinjector. To section V₁, a small cut was made through the skin just above the eye and the eyeball and muscles were gently pushed to the side, so that V₁ could be sectioned and a 2–3-mm piece of the nerve removed. The extent and degree of overlap between cluster N and the cluster N lesions were reconstructed post mortem using AMIRA software (Visage Imaging).

The magnetic compass experiments were performed in Emlen funnels inside four wooden huts lined with grounded aluminium shields to minimize electromagnetic disturbances. Immediately before the orientation tests, we exposed all test birds to parts of the local evening sky.

The sun compass experiments were performed on clear evenings around the time of sunset on an open field. The star compass experiments were performed in a planetarium simulating the local starry sky of Oldenburg. The magnetic field inside the planetarium was strongly disturbed.

Two to three independent observers determined the mean directions in the single tests, and all oriented and active tests were used to calculate the mean orientation of each individual bird under each experimental condition. These individual mean directions are depicted as circles on the peripheries of the large circles in Figs 1–3. The group mean vectors were calculated by vector addition of individual unit vectors in each of the individual bird mean directions and division by the number of birds tested.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 August; accepted 23 September 2009.

1. Wiltschko, W. & Wiltschko, R. Magnetic compass of European robins. *Science* **176**, 62–64 (1972).
2. Wiltschko, W. & Wiltschko, R. Magnetic compass orientation in birds and its physiological basis. *Naturwissenschaften* **89**, 445–452 (2002).
3. Cochran, W. W., Mouritsen, H. & Wikelski, M. Free-flying migrating songbirds recalibrate their magnetic compass daily from sunset cues. *Science* **304**, 405–408 (2004).
4. Mouritsen, H. & Ritz, T. Magnetoreception and its use in bird navigation. *Curr. Opin. Neurobiol.* **15**, 406–414 (2005).
5. Kirschvink, J. L. & Gould, J. L. Biogenic magnetite as a basis for magnetic field sensitivity in animals. *Biosystems* **13**, 181–201 (1981).
6. Walker, M. M. *et al.* Structure and function of the vertebrate magnetic sense. *Nature* **390**, 371–376 (1997).

7. Williams, N. M. & Wild, J. M. Trigeminally innervated iron-containing structures in the beak of homing pigeons, and other birds. *Brain Res.* **889**, 243–246 (2001).
8. Fleissner, G. *et al.* Ultrastructural analysis of a putative magnetoreceptor in the beak of homing pigeons. *J. Comp. Neurol.* **458**, 350–360 (2003).
9. Fleissner, G., Stahl, B., Thalau, P., Falkenberg, G. & Fleissner, G. A novel concept of Fe-mineral-based magnetoreception: histological and physicochemical data from the upper beak of homing pigeons. *Naturwissenschaften* **94**, 631–642 (2007).
10. Mora, C. V., Davison, M., Wild, J. M. & Walker, M. M. Magnetoreception and its trigeminal mediation in the homing pigeon. *Nature* **432**, 508–511 (2004).
11. Ritz, T., Adem, S. & Schulten, K. A model for photoreceptor-based magnetoreception in birds. *Biophys. J.* **78**, 707–718 (2000).
12. Ritz, T., Thalau, P., Phillips, J. B., Wiltschko, R. & Wiltschko, W. Resonance effects indicate radical pair mechanism for avian magnetic compass. *Nature* **429**, 177–180 (2004).
13. Mouritsen, H. *et al.* Cryptochromes and activity markers co-localize in bird retina during magnetic orientation. *Proc. Natl Acad. Sci. USA* **101**, 14294–14299 (2004).
14. Liedvogel, M. *et al.* Chemical magnetoreception: bird cryptochrome 1a is excited by blue light and forms long-lived radical-pairs. *PLoS ONE* **2**, e1106 (2007).
15. Ritz, T. *et al.* Magnetic compass of birds is based on a molecule with optimal directional sensitivity. *Biophys. J.* **96**, 3451–3457 (2009).
16. Mouritsen, H., Feenders, G., Liedvogel, M., Wada, K. & Jarvis, E. D. A night vision brain area in migratory songbirds. *Proc. Natl Acad. Sci. USA* **102**, 8339–8344 (2005).
17. Liedvogel, M. *et al.* Lateralised activation of cluster N in the brains of migratory songbirds. *Eur. J. Neurosci.* **25**, 1166–1173 (2007).
18. Heyers, D., Manns, M., Luksch, H., Güntürkün, O. & Mouritsen, H. A visual pathway links brain structures active during magnetic compass orientation in migratory birds. *PLoS ONE* **2**, e937 (2007).
19. Feenders, G. *et al.* Molecular mapping of movement-associated areas in the avian brain: a motor theory for vocal learning origin. *PLoS ONE* **3**, e1768 (2008).
20. Emlen, S. T. & Emlen, J. T. A technique for recording migratory orientation of captive birds. *Auk* **83**, 361–367 (1966).
21. Mouritsen, H., Feenders, G., Hegemann, A. & Liedvogel, M. Thermal paper can replace typewriter correction paper in Emlen funnels. *J. Ornithol.* **150**, 713–715 (2009).
22. Barami, K., Iversen, K., Furneaux, H. & Goldman, S. A. Hu protein as an early marker of neuronal phenotypic differentiation by subependymal zone cells of the adult songbird forebrain. *J. Neurobiol.* **28**, 82–101 (1995).
23. Pasternak, T. & Hodos, W. Intensity difference thresholds after lesions of the visual Wulst in pigeons. *J. Comp. Physiol. Psychol.* **91**, 485–497 (1977).
24. Åkesson, S. & Sandberg, R. Migratory orientation of passerines at dusk, night and dawn. *Ethology* **98**, 177–191 (1994).
25. Beason, R. & Semm, P. Does the avian ophthalmic nerve carry magnetic navigational information? *J. Exp. Biol.* **199**, 1241–1244 (1996).
26. Presti, D. & Pettigrew, J. D. Ferromagnetic coupling to muscle receptors as a basis for geomagnetic field sensitivity in animals. *Nature* **285**, 99–101 (1980).
27. Dennis, T. E., Raynor, M. J. & Walker, M. M. Evidence that pigeons orient to geomagnetic intensity during homing. *Proc. R. Soc. B* **274**, 1153–1158 (2007).
28. Gagliardo, A., Ioale, P., Savini, M. & Wild, J. M. Having the nerve to home: olfactory versus magnetoreceptor mediation of homing in pigeons. *J. Exp. Biol.* **209**, 2888–2892 (2006).
29. Gagliardo, A., Ioale, P., Savini, M. & Wild, J. M. Navigational abilities of homing pigeons deprived of olfactory or trigeminally mediated magnetic information when young. *J. Exp. Biol.* **211**, 2046–2051 (2008).
30. Kirschvink, J. L. Uniform magnetic fields and double-wrapped coil systems: improved techniques for the design of bioelectromagnetic experiments. *Bioelectromagnetics* **13**, 401–411 (1992).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Bourdonnais, D. Hugo, A. Kittel, C. Mora and several volunteer students for assistance, E. Jarvis for scientific discussions, Blumberg GmbH, Rattigen, Germany for providing the thermal paper, the workshops of the University of Oldenburg for building huts, magnetic coil systems and electronic controls and J. Rahn for assistance in the planetarium of the Fachhochschule Oldenburg/Elsfleth. Financial support was provided by the Volkswagenstiftung (to H.M. and D.H.) and by the Deutsche Forschungsgemeinschaft (to H.M.).

Author Contributions H.M. designed and supervised the study. M.Z., C.M.H., S.E. and J.H. performed and M.Z. and C.M.H. supervised the majority of the orientation experiments. M.Z., C.M.H., S.E., J.H. and H.M. analysed the orientation results. J.M.W. and D.H. performed the surgeries. D.H. did the post-mortem histological analyses. D.D. performed the lesion analyses using AMIRA. S.W. and D.K. performed and analysed the operant conditioning. N.-L.S. suggested and made crucial improvements to the experimental set-up. H.M., M.Z., J.M.W. and D.H. wrote most of the paper. All authors read and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.M. (henrik.mouritsen@uni-oldenburg.de).

METHODS

Magnetic fields. To produce the CMF condition, current flowed in the same direction through each subset of windings of the double-wrapped, three-dimensional Merritt four-coil systems³⁰. To produce the natural geomagnetic field condition, the coils were turned on but the current ran through the two subsets of windings in opposite directions. Before each experiment, the actual magnetic field was measured (FVM-400, Meda) in the centre and at the edges of the experimental volume, within which nine Emlen funnels were placed simultaneously. The actual fields experienced by the birds under the three magnetic field conditions were as follows (mean \pm s.d.): for the NMF the field strength was $48,620 \pm 330$ nT, the inclination was $67.6^\circ \pm 0.5^\circ$ and the horizontal direction was $360^\circ \pm 1^\circ$; for the CMF the field strength was $48,800 \pm 320$ nT, the inclination was $67.8^\circ \pm 0.5^\circ$ and the horizontal direction was $240^\circ \pm 2^\circ$; and for the IMF the field strength was $48,660 \pm 500$ nT, the inclination was $-68.1^\circ \pm 0.5^\circ$ and the horizontal direction was $1^\circ \pm 1^\circ$.

Cluster N lesioning and nerve sectioning. Birds were anaesthetized by intramuscular injection of ketamine (Pfizer)/Rompun (Bayer). The heads were then fixed in a custom-built stereotaxic apparatus. The scalp was additionally anaesthetized using a local surface anaesthetic (xylocaine, Astra Zeneca), incised and retracted. For cluster N lesions, a small part of the skull was carefully removed and 50 nl 1% ibotenic acid in 0.9% NaCl solution were stereotactically injected using a microinjector. The coordinates used corresponded to the ones described in ref. 18. Sham-lesioned birds underwent exactly the same procedures except that ibotenic acid was not injected. For nerve sectioning, the scalp was retracted and the fascia along the rim of the orbit was incised to allow gentle depression and retraction of the globe. The ophthalmic branch of the trigeminal nerve (V_1) was revealed and sectioned behind the eyeball immediately before it left the orbit and 2–3 mm farther proximally; this piece was then removed and the cut ends of the proximal and distal stumps were sealed with surgical cyanoacrylate to prevent re-fusion. Sham-sectioned birds underwent exactly the same procedures except that the nerves were not sectioned. Finally, the skin edges were sealed and the birds were given 1–10 weeks to recover from surgery before taking part in any behavioural experiment.

Behavioural experiments. The experimental birds were caught within 200 m of the testing site at the University of Oldenburg, Germany, during autumn migration and tested during the following spring migratory season. The magnetic field orientation experiments were conducted in wooden huts placed on the university campus. The walls of the huts were lined with grounded aluminium shields, which acted as Faraday cages to shield non-stationary electromagnetic disturbances. One hour (± 10 min) before the lights went out in the bird rooms (light for 14 h, dark for 10 h), the birds were placed outdoors in wooden transport cages that allowed them to see parts of the evening sky for 1 h to give them the opportunity to calibrate their magnetic compass from the local sunset sky³. Immediately thereafter, the birds were placed in aluminium Emlen funnels²⁰ (35 cm in diameter, 15 cm high, walls inclined at 45°) and tested for 1 h under dim light conditions (4 mW m^{-2}) produced by incandescent bulbs (for spectrum, see Supplementary Fig. 2).

The funnels were coated with thermal paper²¹ on which the birds left scratches as they moved. Nine European robins were tested simultaneously in each hut. The birds were put into a randomized funnel position each night, and were put into the funnels from different sides. We observed no systematic differences between the nine funnel positions. The birds were tested twice each night under the same magnetic field condition. There were no statistically significant differences (Mardia–Watson–Wheeler tests, $P > 0.65$ for all comparisons) or even indications that the orientation was systematically different during the first test relative to the second test, and because any given bird was tested in different huts during the first and second tests, both values were entered into the calculation of the mean for that bird under the given magnetic field condition. The first test started 30 min after sunset and the second test started around 2 h after sunset.

The magnetic field condition present in a given hut was switched every second night, and usually different magnetic fields were present in different huts on any given night.

Orientation experiments during sunset took place in spring on an open field near Gristede, 20 km north-northwest of Oldenburg. The tests were performed in the undisturbed local magnetic field with an open view of the sky on clear evenings and the test started 30 min before sunset and ended 30 min after sunset.

Orientation experiments in the planetarium of Elsflath, 28 km east-northeast of Oldenburg, were conducted under a stationary simulation of the local sky. The starry sky was projected using a Zeiss ZKP-2 projector on a dome (9 m in diameter). Nine Emlen funnels were placed symmetrically around the projector 10 cm above the horizon plane, such that the projector was never visible from inside any of the funnels. To increase the likelihood that the magnetic field of the planetarium (already disturbed by the iron-containing projector and magnetic material in the walls) would not provide any useful magnetic compass information, heavy iron racks were used for stabilization of the set-up and strong neodymium magnets were attached under each funnel. The resulting magnetic field inside each funnel varied strongly in intensity (from $\sim 44,000$ nT to $\sim 84,000$ nT), direction (owing to the central placement of the magnet, the horizontal direction of the field within the funnels varied in all directions depending on exactly where the bird was located within the funnel) and inclination (from $+59^\circ$ to $+88^\circ$). Other testing procedures were the same as in the magnetic field orientation experiments, except that the birds were only tested once per night. Before each testing session, the projector was adjusted to the time the birds were tested, and on every second day, the star pattern was adjusted to the actual date.

Orientation-data analysis. Two independent researchers, who did not know either the test condition or the operation a given bird had experienced, determined each bird's mean direction from the distribution of the scratches. If both observers considered the scratches to be randomly distributed or if the two mean directions deviated by more than 30° , a third independent researcher determined the mean direction. If this third individual determined a mean direction similar to one of the first two, and if the individual with the initially differing opinion also agreed with this direction, the mean of the two similar directions was recorded as the orientation result. If the three independent individuals could not agree on one mean direction, the bird's heading was defined as random and excluded from the analyses (only 10% of all tests were excluded on the basis of this criterion). Birds with fewer than 35 scratches on the paper were considered inactive and also excluded from the analysis (the birds were inactive in 18% of all tests).

Size and position of lesion analysis. To determine the exact extent and location of the lesions relative to cluster N, each brain slice was photographed with a digital camera (Leica DFC 320) through a stereo microscope (Leica M, Leica IM50). On these pictures, boundaries of the whole telencephalon, the injured tissue (lesion) and cluster N were marked, and the sections were aligned using Photoshop 6.0/Illustrator 10.0 (Adobe Systems).

The extent of cluster N was determined by comparison with ZENK in-situ hybridized brain slices from birds performing magnetic compass orientation in a funnel^{16,17}. The stacks were aligned using the outline of the telencephalon. Stacks of each hemisphere were launched in AMIRA (Visage Imaging) and converted into AMIRA files (AMIRA mesh binary). The resolution of AMIRA files was reduced to 800×600 pixels and the physical distance between slides—the actual distance between each slide of each series—was set at $240 \mu\text{m}$. The file sequence was fused into one data stack. 'Label fields' were created, in which the marked boundaries were labelled and interpolated into three three-dimensional bodies: the lesion, cluster N and the telencephalon. On the basis of the overlap in space between these three volumes, the percentage of the volumetric overlap between the lesion and cluster N was determined. This procedure was done for each hemisphere of each bird separately.

LETTERS

Regulation of cortical microcircuits by unitary GABA-mediated volume transmission

Szabolcs Oláh¹, Miklós Füle¹, Gergely Komlósi¹, Csaba Varga¹, Rita Báldi¹, Pál Barzó² & Gábor Tamás¹

GABA (γ -aminobutyric acid) is predominantly released by local interneurons in the cerebral cortex to particular subcellular domains of the target cells^{1,2}. This suggests that compartmentalized, synapse-specific action of GABA is required in cortical networks for phasic inhibition^{2–4}. However, GABA released at the synaptic cleft diffuses to receptors outside the postsynaptic density and thus tonically activates extrasynaptic GABA_A and GABA_B receptors, which include subtypes of both receptor families especially sensitive to low concentrations of GABA^{3–7}. The synaptic and extrasynaptic action of GABA corroborates the idea that neurons of the brain use synaptic (or wiring) transmission and non-synaptic (or volume) transmission for communication^{8,9}. However, re-uptake mechanisms restrict the spatial extent of extrasynaptic GABA-mediated effects^{10,11}, and it has been proposed that the concerted action of several presynaptic interneurons, the sustained firing of individual cells or an increase in release-site density is required to reach ambient GABA levels sufficient to activate extrasynaptic receptors^{4,9,11–13}. Here we show that individual neurogliaform cells release enough GABA for volume transmission within the axonal cloud and, thus, that neurogliaform cells do not require synapses to produce inhibitory responses in the overwhelming majority of nearby neurons. Neurogliaform cells suppress connections between other neurons acting on presynaptic terminals that do not receive synapses at all in the cerebral cortex. They also reach extrasynaptic, δ -subunit-containing GABA_A (GABA_{A δ}) receptors responsible for tonic inhibition. We show that GABA_{A δ} receptors are localized to neurogliaform cells preferentially among cortical interneurons. Neurosteroids, which are modulators of GABA_{A δ} receptors, alter unitary GABA-mediated effects between neurogliaform cells. In contrast to the specifically placed synapses formed by other interneurons, the output of neurosteroid-sensitive neurogliaform cells represents the ultimate form of the lack of spatial specificity in GABA-mediated systems, leading to long-lasting network hyperpolarization combined with widespread suppression of communication in the local circuit.

Uniquely among neocortical interneurons, neurogliaform cells evoke long-lasting inhibition in the form of an unusually slow GABA_A-receptor-mediated component and slow GABA_B-receptor-mediated responses in their target neurons^{14–16}. A distinctive feature of neurogliaform cells among cortical interneurons is that they have very dense axonal arborizations, in which presynaptic boutons on the same or neighbouring axon collaterals can be found a couple of micrometres from each other^{14,17} (Fig. 1). GABA can activate receptors located up to several micrometres from the release site¹¹. We measured the density of terminals in neurogliaform-cell ($n = 8$) and basket-cell ($n = 6$) axons ($421,213 \pm 34,289$ and $78,506 \pm 8,423$ boutons per cubic millimetre, respectively; $P < 0.0001$) and found that a single neurogliaform-cell axon matches the potential release-site density of

five or six overlapping basket-cell axons. We proposed that the high density of neurogliaform-cell axons could counteract transmitter re-uptake mechanisms and that GABA released from neurogliaform cells acts as a volume transmitter to reach receptors at synaptic and non-synaptic sites in the tissue the axon intermingles with.

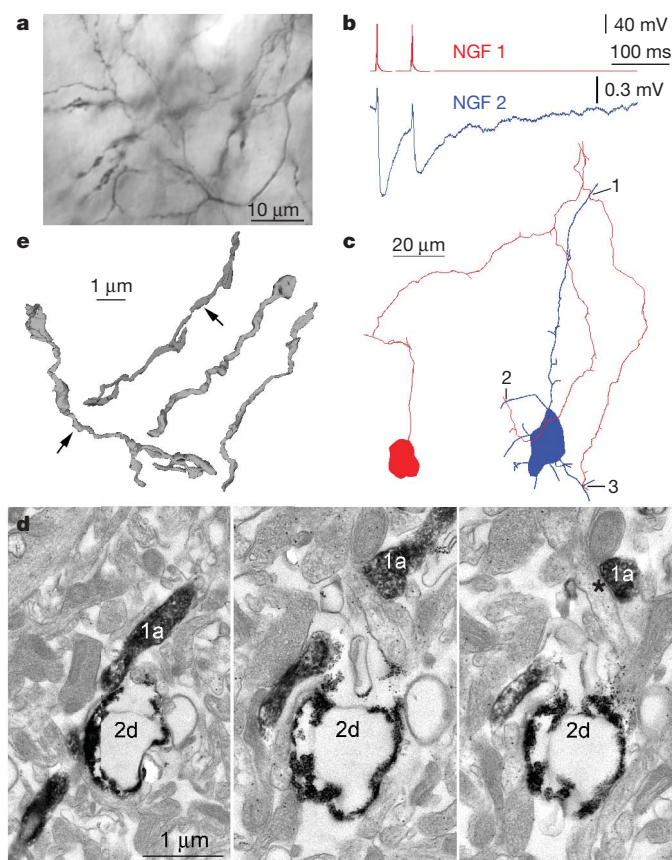


Figure 1 | Neurogliaform cells do not require direct synaptic junctions to affect target cells. **a**, Dense axonal cloud formed by a single neurogliaform cell. **b**, Action potentials in neurogliaform cell (NGF) 1 elicited electrical coupling potentials combined with inhibitory postsynaptic potentials (IPSPs) in NGF 2. **c**, Route of the axon of NGF 1 (red) to close appositions (labelled 1–3) with the dendrites of NGF 2 (blue). None of these appositions could be verified as a synaptic junction. **d**, A non-synaptic close apposition (labelled 1 in **c**) in consecutive serial ultrathin sections (1a, axon of NGF 1; 2d, dendrite of NGF 2). The axon of NGF 1 forms a synaptic junction (asterisk) on an unlabelled dendritic shaft. **e**, Three-dimensional electron microscopic reconstructions of 18 neurogliaform-cell axonal varicosities, of which only two formed synaptic junctions (arrows).

¹Research Group for Cortical Microcircuits of the Hungarian Academy of Sciences, Department of Physiology, Anatomy and Neuroscience, University of Szeged, Közép fasor 52, Szeged H-6726, Hungary. ²Department of Neurosurgery, University of Szeged, Semmelweis utca 6, Szeged H-6725, Hungary.

Potential volume transmission suggests a very high rate of functional coupling between neurogliaform cells and neighbouring neurons. When searching our database, which contains information on 204 simultaneously recorded pairs of neurogliaform cells and other neurons with somata located $<100\mu\text{m}$ apart, we detected hyperpolarizing effects of neurogliaform cells in 178 (87%) of tested cells, an unprecedentedly high proportion in paired recordings of cortical neurons¹⁸. In searching for the morphological correlates of these connections, correlated light and electron microscopic analysis was performed on $n = 11$ neurogliaform-cell-to-interneuron pairs (eight from rat, of which three featured in ref. 19, and three from human) and $n = 5$ neurogliaform-cell-to-pyramidal-cell pairs (rat).

We detected chemical synaptic junctions in only two neurogliaform-to-pyramidal pairs, supporting earlier results¹⁴, but in assessing fully available series of ultrathin sections by tracing neurogliaform-cell axons along their approach to functionally coupled neurons, we did not find synaptic junctions in the remaining 14 cell pairs. In these 14 pairs, the most closely placed synapses established by the boutons of the neurogliaform cells were $2.7 \pm 1.6\mu\text{m}$ from the target dendrites (range, $1.1\text{--}5.3\mu\text{m}$; Fig. 1). Proving the efficacy of our analysis, we confirmed the presence of gap junctions in the three pairs¹⁹ in which electrical coupling potentials were recorded in addition to IPSPs, even though ultrastructural identification of gap junctions between labelled neurons is more difficult than that of synapses (Fig. 1). Furthermore, we performed three-dimensional electron microscopic reconstructions of randomly chosen segments of neurogliaform-cell axons. All examined axonal boutons contained synaptic vesicles, but we found that 50 boutons formed only 11 synapses, implying that the majority ($\sim 78\%$) of neurogliaform-cell axonal varicosities do not form classical synapses (Fig. 1). These results suggest that neurogliaform-cell axons do not necessarily require a synaptic contact to elicit inhibitory responses in target cells.

Although suspected non-synaptic communication by neurogliaform cells is consistent with the lack of detectable synapses, positive evidence is required to prove the volume transmission hypothesis. Functional testing of single-cell-initiated non-synaptic signalling works best on potential targets not contacted by synapses at all. Thus, we asked whether GABA released from neurogliaform cells modulates axon terminals that are not targeted in synaptic junctions in the cerebral cortex²⁰ but frequently express GABA_B receptors^{13,16,21–23}. We confirmed that neocortical neurogliaform cells modulate their own axon terminals by means of GABA_B receptors similar to hippocampal neurogliaform cells¹⁶ (Supplementary Fig. 1 and Supplementary Data), but the modulatory action of neurogliaform cells was not limited to homosynaptic silencing of axon terminals.

Heterosynaptic or paracrine effects of neurogliaform cells on axons of other neurons were also suggested by experiments in which we simultaneously recorded from three neurons, namely a pyramidal-cell-to-interneuron connection (test excitatory postsynaptic potentials (EPSPs)) and a neighbouring neurogliaform cell activated 60 ms before the first EPSP and 120 ms before the second (Fig. 2). Alternating trials with and without the activation of neurogliaform cells were performed at a very low frequency (once every 2 min) to avoid activity-dependent loss of neurogliaform-cell output^{14–16}. As expected from the high rate of coupling of neurogliaform cells to other neurons, the decay phase of the hyperpolarizing effect of the neurogliaform cells overlapped with the test EPSPs; the corresponding input resistance changes of the neurons postsynaptic to the EPSPs were $10 \pm 5\%$ and $4 \pm 5\%$, measured 60 and, respectively, 120 ms after the spike in the neurogliaform cells. Accordingly, we corrected the amplitudes of the EPSPs and IPSPs reported below using the corresponding changes in input resistance and driving force.

Switching on the action potential in the neurogliaform cells 60 ms before the spike in the pyramidal cell did not change the amplitude of the first test EPSPs ($n = 5$; $98 \pm 4\%$) relative to control, that is, when the spike in the neurogliaform cell was not elicited (Fig. 2). This indicates that tonic inhibition through GABA_A receptors potentially

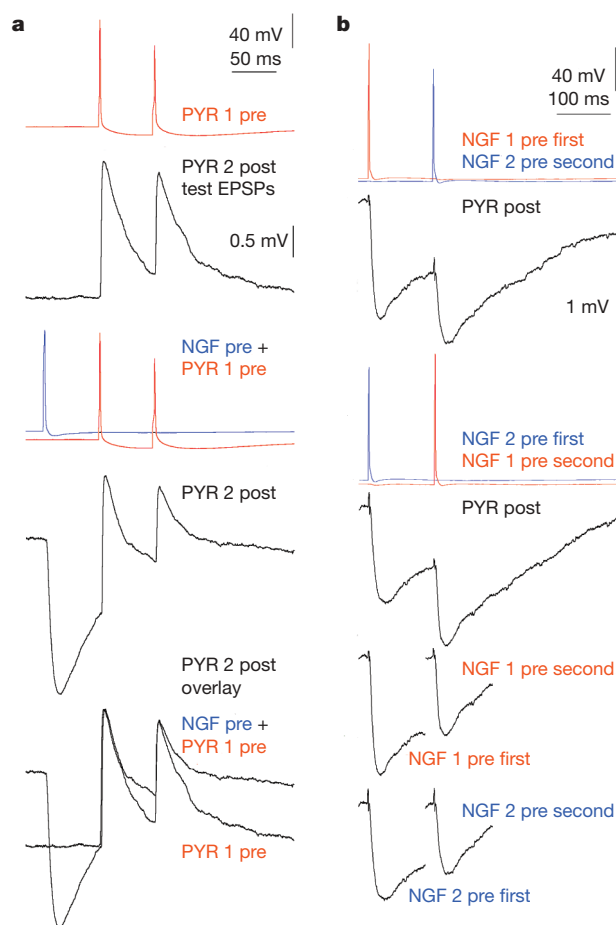


Figure 2 | Heterosynaptic or paracrine effects of neurogliaform cells on axons of other neurons. **a**, Single neurogliaform cells (NGFs) heterosynaptically modulate unitary glutamatergic connections linking other neurons. These simultaneous triple records show a pyramidal-cell-to-pyramidal-cell (PYR 1-to-PYR 2) connection (test EPSPs) while the output of a neurogliaform cell is switched on and off 60 ms before the first pyramidal spike. The neurogliaform cell suppressed the amplitude of the second EPSP evoked by PYR 1, but not the first. **b**, Activation-sequence-dependent cross-modulation of unitary IPSP amplitudes between closely located neurogliaform cells. Top: activation of NGF 1 or NGF 2 followed by a spike in NGF 2 or, respectively, NGF 1 resulted in sequential IPSPs in the postsynaptic pyramidal cell. Bottom: comparison of the amplitude of preceding and succeeding IPSPs elicited by NGF 1 or NGF 2 indicates effective suppression of follower IPSPs.

activated by neurogliaform cells^{15,24} did not interfere significantly with test EPSPs other than by contributing to input resistance changes. However, the neurogliaform cells were effective in decreasing the amplitude of the second test EPSPs, timed 120 ms after the spike in the neurogliaform cells, to $77 \pm 5\%$ ($P < 0.03$) of control (Fig. 2). The differential action of neurogliaform-cell output on the first and the second test EPSPs induced a change in their paired pulse ratio (from $85 \pm 20\%$ to $69 \pm 22\%$; $P < 0.01$). Moreover, neurogliaform cells activated 120 ms before test IPSPs triggered by other neurogliaform cells successfully suppressed the amplitude of the test IPSPs to $74 \pm 4\%$ ($n = 10$, $P < 0.02$) of control (Fig. 2). This resulted in an unusual activation-sequence-dependent cross-modulation of IPSP amplitudes between closely located neurogliaform cells.

Thus, the experiments on test EPSPs and IPSPs showed that individual action potentials in neurogliaform cells could suppress appropriately delayed responses elicited by other neurons. These effects were enhanced by NO-711, a GABA re-uptake blocker (Supplementary Fig. 2 and Supplementary Data). We also asked if basket cells could modulate the surrounding microcircuit similarly to neurogliaform

cells, but single or multiple spikes in these interneurons failed to modulate the amplitude of test connections (Supplementary Fig. 3 and Supplementary Data). Furthermore, we found that GABA_B receptors located on presynaptic terminals²⁵ are necessary and sufficient for the neurogliaform-cell-elicited heterosynaptic suppression of test connections, and supplementary anatomical analysis showed that the modulatory effect of GABA released from neurogliaform cells appears to be confined to the axonal cloud (Supplementary Fig. 4 and Supplementary Data).

Apart from presynaptic GABA_B receptors, extrasynaptically placed GABA_A receptors^{3–7} are potential targets for non-synaptically acting GABA released by neurogliaform cells. Tonic inhibition mediated by GABA_A receptors is variably present on cortical neurons²⁶. We therefore applied immunocytochemistry, which showed that among cortical interneurons, neurogliaform cells are primary candidates for GABA_A-receptor expression (Supplementary Figs 5 and 6 and Supplementary Data). Immunolabelling of electrophysiologically and anatomically identified cells confirmed that neurogliaform cells somatodendritically express GABA_A receptors ($n = 9$) and that some ($n = 3$) of the neurogliaform cells that did so elicited slow IPSPs in simultaneously recorded GABA_A-receptor-immunonegative interneurons (Fig. 3).

Tonic inhibition mediated by GABA_A receptors seems to be a target for stress-induced and ovarian-steroid-derived neurosteroids^{3,4,26,27}; thus, these compounds should modulate the excitability of neurogliaform cells. The average holding current necessary to clamp neurogliaform cells ($n = 7$) but not other interneurons ($n = 10$; Supplementary Fig. 7; ref. 26) at a given holding potential increased by 24 ± 5 pA ($P < 0.03$) after the addition of the neurosteroid tetrahydrodeoxycorticosterone (THDOC; 100 nM)^{3,4,26,27} during blockade of GABA_B receptors with 40 μ M CGP35348 in the presence of 5 μ M GABA (Fig. 3). The effect of THDOC was reversed during blockade of GABA_A receptors with gabazine (10 μ M): average holding currents decreased by 8 ± 3 pA ($P < 0.05$) after addition of THDOC. Furthermore, rheobasic firing of neurogliaform cells ($n = 6$; Fig. 3) but not other interneurons ($n = 8$; Supplementary Fig. 7) required larger positive current injections (234 ± 26 pA) in the presence of THDOC (20 nM), relative to baseline conditions (138 ± 17 pA; $P < 0.01$; 40 μ M CGP35348, 5 μ M GABA), and the effect was abolished with gabazine (10 μ M), indicating that GABA_A receptors could effectively control the input–output gain^{3,28,29} of neurogliaform cells.

In our test, we used low concentrations of externally added GABA mimicking ambient cortical levels, as in earlier experiments on neurosteroid modulation of tonic inhibition^{3,4,26,27}. However, neurogliaform output could locally produce extracellular GABA concentrations effective on GABA_A receptors without external input. Application of THDOC (100 nM) in the presence of CGP35348 (40 μ M) increased the half-width of gabazine (10 μ M)-sensitive IPSPs elicited by single presynaptic action potentials in reciprocally connected pairs of neurogliaform cells from 48 ± 6 ms to 59 ± 10 ms ($n = 5$, $P < 0.04$; Fig. 3). Neurosteroid modulation of self-inhibition and intercellular inhibition involving neurogliaform cells confirms that neurogliaform cells are both sources and targets of extracellular GABA acting on tonic inhibition.

Unlike other interneuron types that specifically place synapses on particular compartments of postsynaptic cells^{1,2}, neurogliaform cells provide non-synaptic, spatially non-specific input to the entire surface of target cells, complementing conventional synapses¹⁴. Neurogliaform cells release GABA, covering their axonal fields in effective concentrations, and target the overwhelming majority of nearby neurons, which selectively express receptors sensitive to low concentrations of the neurotransmitter on their various compartments^{23,25,30}. Although presynaptic mechanisms producing the GABA cloud around neurogliaform axons are not understood, they might involve a unique release mechanism with an unconventional calcium dependence^{16,22}. Provided that release and re-uptake mechanisms are similar, local GABA concentrations produced by distinct interneurons probably emerge

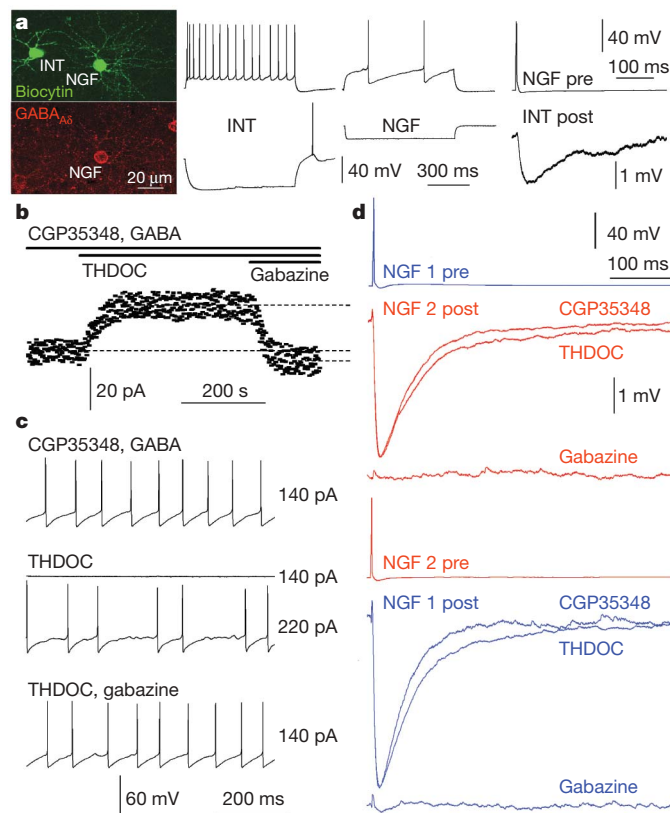


Figure 3 | Extrasynaptically placed GABA_A receptors are localized to neurogliaform cells and targeted by GABA released from neurogliaform cells. **a**, Left: GABA_A-receptor immunoreaction on a simultaneously recorded and biocytin-filled neurogliaform cell (NGF) and postsynaptic interneuron (INT). GABA_A receptors were detected on the neurogliaform cell only. Middle: firing pattern of the interneuron and neurogliaform cell. Right: the neurogliaform cell elicited slow IPSPs in the postsynaptic GABA_A-receptor immunonegative interneuron. **b–d**, Neurosteroids alter the excitability and connections of neurogliaform cells through GABA_A receptors. **b**, The average current (dashed lines) required to hold a neurogliaform cell at the same membrane potential changed after the addition of the neurosteroid THDOC (100 nM) while GABA_B receptors were blocked using 40 μ M CGP35348 in the presence of 5 μ M GABA. When gabazine (10 μ M) was introduced, the effect of THDOC was reversed: the average holding current was smaller than that measured before THDOC addition. **c**, Rheobasic firing of a neurogliaform cell required larger positive current injections (220 pA instead of 140 pA) in the presence of THDOC (20 nM) than under baseline conditions (40 μ M CGP35348, 5 μ M GABA), and the effect was abolished with gabazine (10 μ M). **d**, Application of THDOC (100 nM) in the presence of CGP35348 (40 μ M) increased the half-width of gabazine (10 μ M)-sensitive IPSPs elicited by single presynaptic action potentials in reciprocally connected pairs of neurogliaform cells.

at distances of about half the inter-terminal distance, meaning that basket cells should be less effective than neurogliaform cells around 3 μ m from their terminals¹¹. The spatial extent of axons^{14,17,19} suggests that neurogliaform cells provide a means of making synchronized changes in the efficacy of synaptic connections in conjunction with regulating dendritic excitability across distances of around 200 μ m.

In certain operational states of the microcircuit, solitary spikes in a single neurogliaform cell might replace the concerted action potentials of interneuron populations in modulating presynaptic terminals and postsynaptic domains expressing GABA receptors^{2,4,9,11–13}. Neurosteroids might shift the balance among the sources of ambient GABA by lowering the contribution of neurogliaform cells with a selective increase in tonic inhibition through GABA_A receptors. Varying neurosteroid concentrations during the ovarian cycle and pregnancy and as a result of stress^{4,27} are expected to modulate the action of neurogliaform cells on network hyperpolarization and in

suppressing communication in the local circuit acting on axons of resident neurons or terminals of long-range projections at their arrival.

METHODS SUMMARY

All procedures were performed with the approval of the University of Szeged and in accordance with the US National Institutes of Health Guide to the Care and Use of Laboratory Animals. We obtained electrophysiological data at $\sim 35^\circ\text{C}$ from up to four concomitantly recorded cells visualized in layer 2/3 of the somatosensory cortex of Wistar rats (P22–P35; P, postnatal day) as described previously¹⁴. Presynaptic neurogliaform cells and other cell types were stimulated to elicit action potentials with brief (2-ms) suprathreshold pulses at intervals of $>120\text{ s}$ to avoid exhaustion of transmission. We corrected the amplitudes of postsynaptic potentials riding on top of the decay phase of preceding IPSPs using changes related to input resistance and driving force ($2 \pm 1\%$ for IPSPs and $0.5 \pm 0.2\%$ for EPSPs). Data are given as mean \pm s.d. The Wilcoxon test and Mann–Whitney *U*-test were used to compare data sets. Differences were accepted as significant if $P < 0.05$.

Visualization of biocytin and correlated light and electron microscopy were performed as described previously^{14,18}. Three-dimensional light and electron microscopic reconstructions were carried out using NEUROLUCIDA (MicroBrightfield) and RECONSTRUCT (SYNAPSE WEB) software (version 1.1). Using three-dimensional reconstructions and/or serial ultrathin sections, synaptic junctions were defined as 20–25-nm-wide rigid appositions between pre- and postsynaptic profiles with an accumulation of presynaptic vesicles in axon terminals. The absence of either of these criteria around presynaptic boutons was used to classify a particular axon terminal without synapses.

Immunocytochemistry was performed on adult Wistar rats ($n = 5$) and GABA_A-receptor δ -subunit $-/-$ mice ($n = 2$; donated by I. Mody, University of California, Los Angeles) with standard methods (Methods). For immunoreactions concerning δ -subunits of GABA_A receptors, a primary rabbit polyclonal antibody was used (1:500; gift from W. Sieghart, Center for Brain Research, Vienna, Austria). Co-localizations were performed afterwards with primary antibodies as described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 July; accepted 16 September 2009.

- Freund, T. F. & Buzsáki, G. Interneurons of the hippocampus. *Hippocampus* **6**, 347–470 (1996).
- Klausberger, T. & Somogyi, P. Neuronal diversity and temporal dynamics: the unity of hippocampal circuit operations. *Science* **321**, 53–57 (2008).
- Farrant, M. & Nusser, Z. Variations on an inhibitory theme: phasic and tonic activation of GABA(A) receptors. *Nature Rev. Neurosci.* **6**, 215–229 (2005).
- Glykys, J. & Mody, I. Activation of GABAA receptors: views from outside the synaptic cleft. *Neuron* **56**, 763–770 (2007).
- Nusser, Z., Sieghart, W. & Somogyi, P. Segregation of different GABAA receptors to synaptic and extrasynaptic membranes of cerebellar granule cells. *J. Neurosci.* **18**, 1693–1703 (1998).
- Fritschy, J. M. & Brunig, I. Formation and plasticity of GABAergic synapses: physiological mechanisms and pathophysiological implications. *Pharmacol. Ther.* **98**, 299–323 (2003).
- Moss, S. J. & Smart, T. G. Constructing inhibitory synapses. *Nature Rev. Neurosci.* **2**, 240–250 (2001).
- Vizi, E. S. Role of high-affinity receptors and membrane transporters in nonsynaptic communication and drug action in the central nervous system. *Pharmacol. Rev.* **52**, 63–89 (2000).
- Barbour, B. & Häusser, M. Intersynaptic diffusion of neurotransmitter. *Trends Neurosci.* **20**, 377–384 (1997).
- Guastella, J. *et al.* Cloning and expression of a rat brain GABA transporter. *Science* **249**, 1303–1306 (1990).

- Overstreet, L. S. & Westbrook, G. L. Synapse density regulates independence at unitary inhibitory synapses. *J. Neurosci.* **23**, 2618–2626 (2003).
- Scanziani, M. GABA spillover activates postsynaptic GABA(B) receptors to control rhythmic hippocampal activity. *Neuron* **25**, 673–681 (2000).
- Mitchell, S. J. & Silver, R. A. GABA spillover from single inhibitory axons suppresses low-frequency excitatory transmission at the cerebellar glomerulus. *J. Neurosci.* **20**, 8651–8658 (2000).
- Tamás, G., Lörincz, A., Simon, A. & Szabadics, J. Identified sources and targets of slow inhibition in the neocortex. *Science* **299**, 1902–1905 (2003).
- Szabadics, J., Tamás, G. & Soltesz, I. Different transmitter transients underlie presynaptic cell type specificity of GABA_{A,slow} and GABA_{A,fast}. *Proc. Natl Acad. Sci. USA* **104**, 14831–14836 (2007).
- Price, C. J., Scott, R., Rusakov, D. A. & Capogna, M. GABA(B) receptor modulation of feedforward inhibition through hippocampal neurogliaform cells. *J. Neurosci.* **28**, 6974–6982 (2008).
- Karube, F., Kubota, Y. & Kawaguchi, Y. Axon branching and synaptic bouton phenotypes in GABAergic nonpyramidal cell subtypes. *J. Neurosci.* **24**, 2853–2865 (2004).
- Markram, H. *et al.* Interneurons of the neocortical inhibitory system. *Nature Rev. Neurosci.* **5**, 793–807 (2004).
- Simon, A., Olah, S., Molnar, G., Szabadics, J. & Tamás, G. Gap-junctional coupling between neurogliaform cells and various interneuron types in the neocortex. *J. Neurosci.* **25**, 6278–6285 (2005).
- Peters, A. P., Palay, S. L. & de F. Webster, H. *The Fine Structure of the Nervous System* (Oxford Univ. Press, 1991).
- Isaacson, J. S., Solis, J. M. & Nicoll, R. A. Local and diffuse synaptic actions of GABA in the hippocampus. *Neuron* **10**, 165–175 (1993).
- Sakaba, T. & Neher, E. Direct modulation of synaptic vesicle priming by GABA(B) receptor activation at a glutamatergic synapse. *Nature* **424**, 775–778 (2003).
- Guetg, N. *et al.* The GABAB1a isoform mediates heterosynaptic depression at hippocampal mossy fiber synapses. *J. Neurosci.* **29**, 1414–1423 (2009).
- Pearce, R. A. Physiological evidence for two distinct GABAA responses in rat hippocampus. *Neuron* **10**, 189–200 (1993).
- Kulik, A. *et al.* Subcellular localization of metabotropic GABA(B) receptor subunits GABA(B1a/b) and GABA(B2) in the rat hippocampus. *J. Neurosci.* **23**, 11026–11035 (2003).
- Vardya, I., Drasbek, K. R., Dosa, Z. & Jensen, K. Cell type-specific GABA A receptor-mediated tonic inhibition in mouse neocortex. *J. Neurophysiol.* **100**, 526–532 (2008).
- Stell, B. M., Brickley, S. G., Tang, C. Y., Farrant, M. & Mody, I. Neuroactive steroids reduce neuronal excitability by selectively enhancing tonic inhibition mediated by delta subunit-containing GABAA receptors. *Proc. Natl Acad. Sci. USA* **100**, 14439–14444 (2003).
- Brickley, S. G., Revilla, V., Cull-Candy, S. G., Wisden, W. & Farrant, M. Adaptive regulation of neuronal excitability by a voltage-independent potassium conductance. *Nature* **409**, 88–92 (2001).
- Chadderton, P., Margrie, T. W. & Häusser, M. Integration of quanta in cerebellar granule cells during sensory processing. *Nature* **428**, 856–860 (2004).
- Kullmann, D. M. *et al.* Presynaptic, extrasynaptic and axonal GABAA receptors in the CNS: where and why? *Prog. Biophys. Mol. Biol.* **87**, 33–46 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank W. Sieghart for donating antibody, A. Lörincz and Z. Nusser for initial testing of the GABA_{A δ} antibody, I. Mody for the GABA_{A δ} $-/-$ animals, and A. Simon and E. Tóth for reconstructions. This work was supported by the European Young Investigator Award, the Hungarian National Office for Research and Technology Polányi Award, the Howard Hughes Medical Institute, US National Institutes of Health grant NS535915, the Boehringer Ingelheim Fonds and the Hungarian Academy of Sciences.

Author Contributions S.O. performed experiments, analysed data and wrote the paper; M.F., G.K., C.V. and R.B. performed experiments and analysed data; P.B. performed experiments; and G.T. designed and performed experiments, analysed data and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to G.T. (gtamas@bio.u-szeged.hu).

METHODS

Electrophysiology. All procedures were performed with the approval of the University of Szeged and in accordance with the US National Institutes of Health Guide to the Care and Use of Laboratory Animals. Wistar rats (P22–P35) were anaesthetized by the intraperitoneal injection of ketamine (30 mg kg⁻¹) and xylazine (10 mg kg⁻¹), and following decapitation, coronal slices (350 µm thick) were prepared from the somatosensory cortex. Slices were incubated at room temperature (25 °C) for 1 h in a solution composed of 130 mM NaCl, 3.5 mM KCl, 1 mM NaH₂PO₄, 24 mM NaHCO₃, 1 mM CaCl₂, 3 mM MgSO₄, 10 nM D(+)-glucose, saturated with 95% O₂ and 5% CO₂. The solution used during recordings differed only in that it contained 3 mM CaCl₂ and 1.5 mM MgSO₄. Recordings were obtained at ~35 °C from up to four concomitantly recorded cells visualized in layer 2/3 by infrared differential-interference-contrast video microscopy (Olympus BX60WI microscope, Hamamatsu CCD camera, Luigs & Neumann Infrapatch set-up and two HEKA EPC 10 Double patch-clamp amplifiers). Micropipettes (5–7 MΩ) were filled with 126 mM K gluconate, 4 mM KCl, 4 mM ATP-Mg, 0.3 mM GTP-NA₂, 10 mM HEPES, 10 mM creatine phosphate and 8 mM biocytin (pH 7.25, 300 mosmol). Signals were filtered at 5 kHz, digitized at 10 kHz and analysed with PULSE software (version 8.54; HEKA).

Presynaptic neurogliaform cells and other cell types were stimulated to elicit action potentials with brief (2-ms) suprathreshold pulses at intervals of >120 s to avoid exhaustion of transmission. Postsynaptic cells were held at a membrane potential of -51 ± 4 mV. Unless specified otherwise, traces shown are averages of 30–200 episodes. The amplitude of postsynaptic response was defined as the difference between the peak amplitude and the baseline value measured before onset of the postsynaptic potential in control experiments (that is, when additional preceding spikes in the neurogliaform or fast-spiking basket cells were not elicited). These amplitudes were compared with postsynaptic averages recorded with the preceding presynaptic spikes elicited in the neurogliaform or fast-spiking basket cells. Amplitudes of postsynaptic potentials riding on top of the decay phase of preceding IPSPs were corrected using the input-resistance changes measured with brief hyperpolarizing pulses timed on top of the IPSPs at 60 and 120 ms after the spike in the neurogliaform or fast-spiking basket cells. Data were also corrected for driving-force-related changes, which were $2 \pm 1\%$ for IPSPs and $0.5 \pm 0.2\%$ for EPSPs. Experiments with THDOC were performed in the presence of NBQX (10 µM) and APV (20 µM). Effects between neurogliaform cells could not be adequately fitted with single exponentials; thus, half-widths were compared. Data are given as mean \pm s.d. The Wilcoxon test and Mann–Whitney *U*-test were used to compare data sets; differences were accepted as significant if $P < 0.05$.

Histology. Visualization of biocytin and correlated light and electron microscopy was performed as described previously^{14,18}. Three-dimensional light microscopic reconstructions were carried out using NEUROLUCIDA (version 4.05; MicroBrightfield) with a $\times 100$ objective; measurements of the overlap

between axonal and dendritic fields were helped using NEUROEXPLORER (version 3.23; MicroBrightfield) software. Three-dimensional electron microscopic reconstructions were performed with RECONSTRUCT (SYNAPSE WEB; version 1.1) software axons and were aided by the especially large range ($\pm 80^\circ$) of the goniometer fitted to our electron microscope (Tecnai BioTwin 120). Using three-dimensional reconstructions and/or serial ultrathin sections, synaptic junctions were defined as 20–25-nm-wide rigid appositions between pre- and postsynaptic profiles with an accumulation of presynaptic vesicles in axon terminals. The absence of either of these criteria around presynaptic boutons was used to classify a particular axon terminal without synapses.

Immunocytochemistry. Adult Wistar rats ($n = 5$) and GABA_A-receptor δ -subunit $-/-$ mice ($n = 2$) were perfused with fixative containing 4% paraformaldehyde in 0.1 M phosphate buffer (PB, pH 7.3) for 10 min in deep anaesthesia. For immunoreactions concerning δ -subunits of GABA_A receptors, 60-µm-thick coronal sections were incubated in citrate buffer containing 10 mM citric acid and 0.05% Tween 20 (Sigma) in distilled water (pH 6.0) at 95–100 °C for 10 min. After cooling to room temperature, sections were blocked with normal horse serum (NHS, 10%) in Tris-buffered saline (TBS, pH 7.4) for 1 h and incubated with primary rabbit polyclonal antibody (1:500) diluted in TBS containing 2% NHS and 0.1% TritonX-100 for 72 h at 4 °C. Washes were done between steps with TBS containing 0.05% Tween 20 until incubation in a cocktail containing biotinylated donkey anti-rabbit secondary antibody (1:250; Jackson ImmunoResearch), and thereafter with TBS only. Sections were treated with ABC complex dissolved in TBS (1:100; Vector Laboratories) and GABA_A δ -receptor immunoreactions were visualized using Alexa 488-conjugated tyramide signal amplification kits (Molecular Probes).

Co-localizations were performed afterwards with the following primary antibodies: mouse anti- α -actinin (1:40,000; Sigma), goat anti-parvalbumin (1:5,000; PGV-214, Swant), mouse anti-calbindin (1:100; C8666, Sigma), guinea pig anti-vasoactive intestinal peptide (1:200; T-5030, Peninsula Laboratories), mouse anti-calretinin (1:1,000; 6B3, Swant), rat anti-somatostatin (1:50; MAB354, Chemicon), mouse anti-chicken ovalbumin upstream promoter transcription factor II (1:500; 2ZH7147H, PPMX) and mouse anti-reelin (1:50,000; MAB5366, Chemicon). Primaries were diluted in the cocktail described above and were visualized using the following secondaries: Cy3-conjugated donkey anti-mouse (1:500; Jackson ImmunoResearch), Cy3-conjugated donkey anti-rabbit (1:500; Jackson ImmunoResearch), Cy3-conjugated donkey anti-rat (1:500; Jackson ImmunoResearch), Cy5-conjugated donkey anti-guinea pig (1:500; Jackson ImmunoResearch) and Alexa 350-conjugated donkey anti-goat (1:500; Molecular Probes). Finally, sections were mounted on slides in Vectashield (Vector Laboratories). Images were made using a light microscope (BX60, Olympus) with a $\times 5$ objective or a confocal laser scanning microscope (IX81, Olympus) with a $\times 20$ (numerical aperture, 0.75) or a $\times 40$ (numerical aperture, 1.30) objective. Automated sequential acquisition of multiple channels was used. Z-stack images were made up from 3–9 images in 5–45 µm depth of tissue.

LETTERS

Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43

Kendle M. Maslowski^{1,2,3}, Angelica T. Vieira^{1,4}, Aylwin Ng⁵, Jan Kranich^{1,2}, Frederic Sierro¹, Di Yu¹, Heidi C. Schilter^{1,2,3}, Michael S. Rolph^{1,2}, Fabienne Mackay^{1,6}, David Artis⁷, Ramnik J. Xavier^{5,8}, Mauro M. Teixeira⁴ & Charles R. Mackay^{1,2,3,6}

The immune system responds to pathogens by a variety of pattern recognition molecules such as the Toll-like receptors (TLRs), which promote recognition of dangerous foreign pathogens. However, recent evidence indicates that normal intestinal microbiota might also positively influence immune responses, and protect against the development of inflammatory diseases^{1,2}. One of these elements may be short-chain fatty acids (SCFAs), which are produced by fermentation of dietary fibre by intestinal microbiota. A feature of human ulcerative colitis and other colitic diseases is a change in 'healthy' microbiota such as *Bifidobacterium* and *Bacteriodes*³, and a concurrent reduction in SCFAs⁴. Moreover, increased intake of fermentable dietary fibre, or SCFAs, seems to be clinically beneficial in the treatment of colitis^{5–9}. SCFAs bind the G-protein-coupled receptor 43 (GPR43, also known as FFAR2)^{10,11}, and here we show that SCFA–GPR43 interactions profoundly affect inflammatory responses. Stimulation of GPR43 by SCFAs was necessary for the normal resolution of certain inflammatory responses, because GPR43-deficient (*Gpr43*^{−/−}) mice showed exacerbated or unresolving inflammation in models of colitis, arthritis and asthma. This seemed to relate to increased production of inflammatory mediators by *Gpr43*^{−/−} immune cells, and increased immune cell recruitment. Germ-free mice, which are devoid of bacteria and express little or no SCFAs, showed a similar dysregulation of certain inflammatory responses. GPR43 binding of SCFAs potentially provides a molecular link between diet, gastrointestinal bacterial metabolism, and immune and inflammatory responses.

Recent evidence suggests that products of intestinal microbiota might positively influence inflammatory disease pathogenesis^{1,2}. To identify factors produced by bacteria that might be beneficial to host immune responses, we induced colitis chemically by adding dextran sulphate sodium (DSS) to the drinking water of germ-free mice. The absence of bacteria from these mice, and the consequences this had on immune responses, resulted in significantly worse colonic inflammation, compared to conventionally raised (CNV) mice. Germ-free mice treated with DSS had decreased weight, an increased daily activity index (DAI; a combined measure of weight loss, rectal bleeding and stool consistency) and a decreased haematocrit (Fig. 1a). Germ-free mice re-colonized with gut microbiota, by gavaging with CNV faeces, showed a marked reduction in inflammation (Supplementary Fig. 1), indicating that the exacerbated response related to lack of bacterial colonisation of the gut. Products of bacteria that have been reported to show anti-inflammatory properties include SCFAs, which are produced by colonic bacteria after fermentation of dietary fibre. Bacteria of the *Bacteroidetes* phylum produce high levels of acetate and propionate, whereas bacteria of the

Firmicutes phylum produce high amounts of butyrate¹². Germ-free mice do not produce SCFAs owing to a lack of enteric microbes¹³. Treatment of germ-free mice with 150 mM acetate in the drinking water markedly improved disease indices, with an increase in colon length, decreased DAI and levels of the inflammatory mediator myeloperoxidase (MPO)

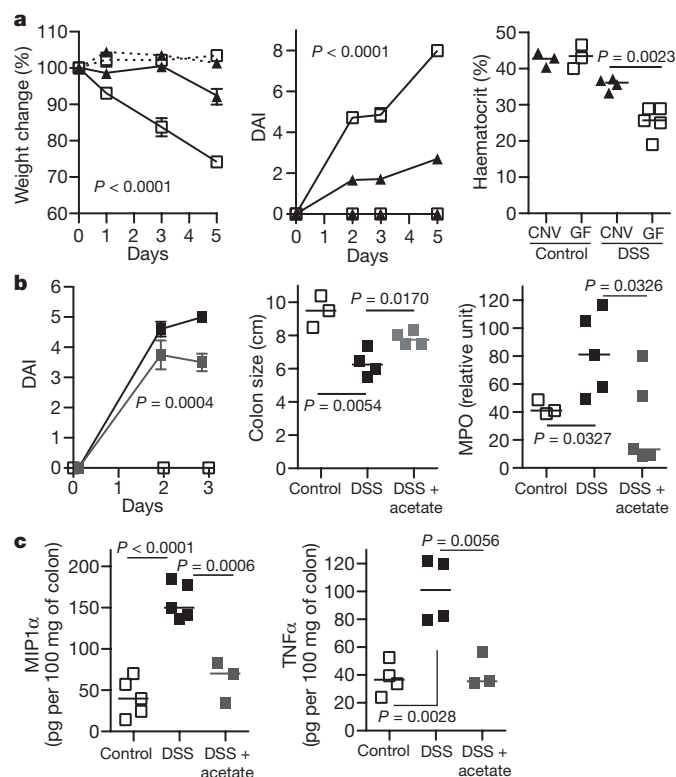


Figure 1 | Exacerbated colitis in germ-free mice is ameliorated by acetate.

a, Germ-free (open squares) and CNV (closed triangles) mice were given DSS colitis (4%), $n = 7$ (experimental groups). Dashed lines, control mice; solid lines, DSS-treated mice. The percentage weight change (left), DAI (middle) and haematocrit (right) were measured. **b**, Germ-free mice were fed acetate (grey squares; 150 mM; $n = 3$) in the drinking water or water only (black squares; $n = 5$), 5 days before and during DSS administration. Control fed denotes no DSS (open squares). Daily activity score (left), colon length (middle) and colonic MPO (right) were determined. **c**, MIP1 α and TNF α levels in acetate-fed mice. Data are median \pm s.e.m., representative of two independent experiments.

¹Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, New South Wales 2010, Australia. ²Cooperative Research Center for Asthma and Airways, Camperdown, New South Wales 2050, Australia. ³St Vincent's Clinical School, University of New South Wales, New South Wales 2010, Australia. ⁴Department of Biochemistry and Immunology, Instituto de Ciencias Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais 31270-901, Brazil. ⁵Center for Computational and Integrative Biology and Gastrointestinal Unit, Massachusetts General Hospital and Harvard Medical School, 185 Cambridge Street, Boston, Massachusetts 02114, USA. ⁶Faculty of Medicine, Monash University, Wellington Road, Clayton, Victoria 3800, Australia. ⁷School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. ⁸Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.

(Fig. 1b), and decreased levels of $\text{TNF}\alpha$ and inflammatory MIP1 α (also known as CCL3) (Fig. 1c). SCFAs have a well-characterized anti-inflammatory effect, on both colonic epithelium and immune cells^{14–17}. Recently, SCFAs, particularly acetate (C2) and propionate (C3), have been found to bind and activate the G-protein-coupled receptor GPR43 (refs 10, 11). Using an extensive data set of human and mouse immune cell transcription profiles¹⁸ we found that transcripts for human *GPR43* and mouse *Gpr43* exhibited enhanced expression in neutrophils and eosinophils (Fig. 2a and Supplementary Fig. 2). Using nearest-neighbour

correlation analysis we found that *GPR43* gene expression was closely regulated with receptors important for innate immunity, such as Toll-like receptors (TLR2 and TLR4), formyl peptide receptors (FPR1 and FPR2), IL8RB (also known as CXCR2) and C5aR (Fig. 2a). We constructed protein interaction networks for genes co-regulated with GPR43, and, using a molecular complex detection and a graph-theory-based clustering (MCODE) algorithm¹⁹, we identified densely connected local sub-network clusters, revealing modules associated with apoptosis and innate immunity-related processes (Supplementary Fig. 3).

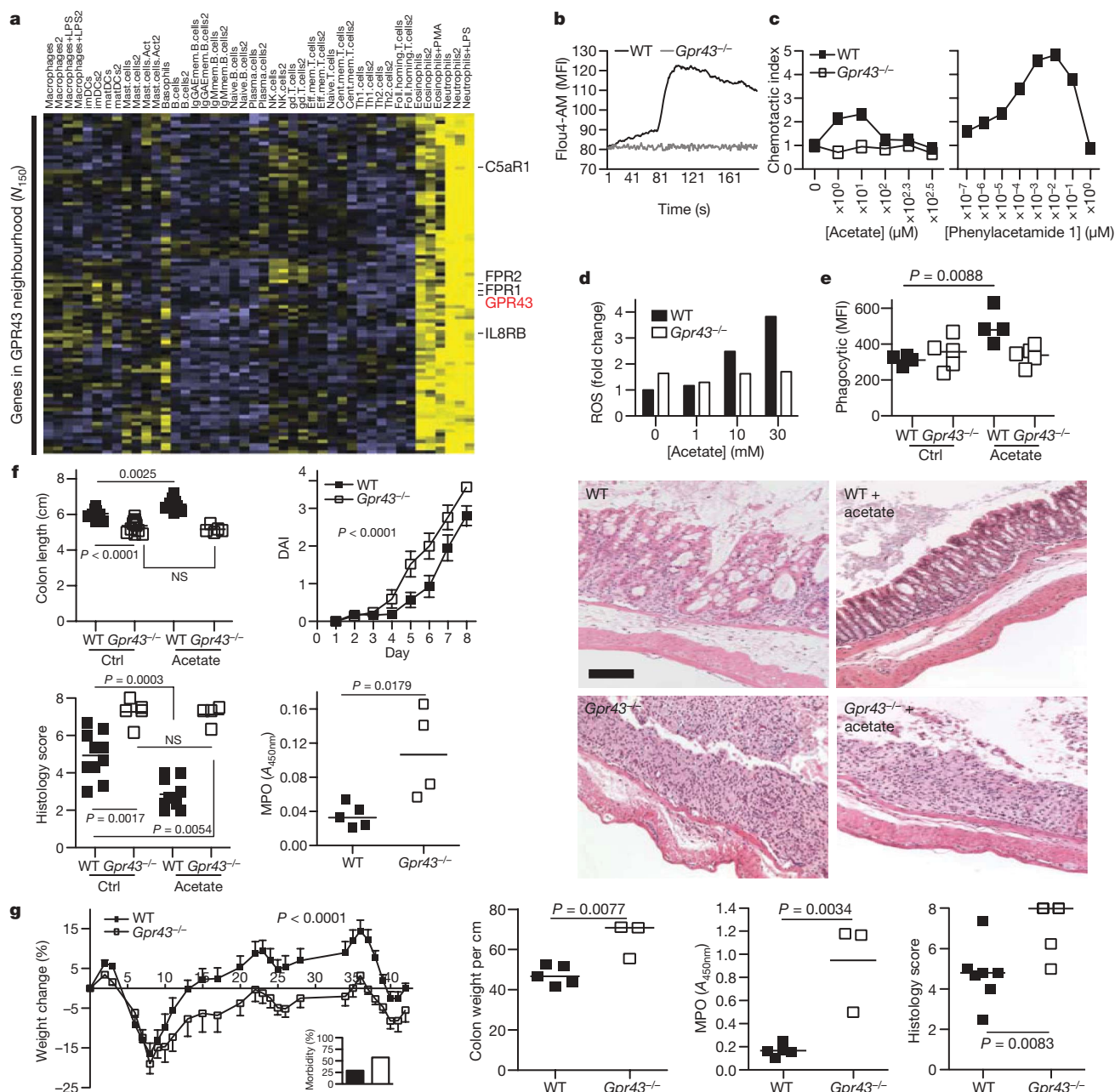


Figure 2 | GPR43 expression and role in inflammatory responses.

a, Immune expression signature of genes encoding cellular receptors across a large panel of leukocyte subsets. Clustering of receptor genes exhibiting enriched expression in neutrophils and eosinophils reveals GPR43, along with other receptors important for innate immunity and chemoattractant-induced responses. Correlation analysis across a wider set of genes in this immune panel identified a rank-ordered list of the top 150 genes (N_{150}) in the co-expression neighbourhood of GPR43. **b–e**, Comparison of wild-type (WT) and *Gpr43*^{−/−} bone marrow neutrophils with respect to acetate-induced Ca^{2+} flux (**b**; MFI, mean fluorescence intensity), chemotaxis (**c**, left panel), ROS production (**d**), and phagocytosis of fluorescently labelled

S. aureus (**e**). The right panel of **c** shows the GPR43 synthetic agonist phenylacetamide 1 in human neutrophil chemotaxis. **f**, DSS colitis (2.5% (w/v)) in wild-type and *Gpr43*^{−/−} mice, fed with acetate or control water (Ctrl). Shown are colon length, the DAI, histology score and colon MPO levels. NS, not significant. The far right panels show representative histological sections from wild-type or *Gpr43*^{−/−} mice as indicated (scale bar, 50 μm). **g**, Chronic DSS-induced colitis ($n = 7$ per group, median \pm s.e.m.). The inset shows the percentage morbidity. Shown are the percentage change in weight, colon weight per cm of colon, colonic MPO, and histological score, for wild-type and *Gpr43*^{−/−} mice.

To demonstrate that GPR43 was the relevant receptor for SCFA effects on immune cells, we sourced *Gpr43*^{-/-} mice (Supplementary Fig. 4), and found that T- and B-cell numbers were normal, and their blood neutrophil numbers were in the normal range (not shown). Acetate induced a robust calcium flux in mouse neutrophils (as well as in human neutrophils and eosinophils, not shown), but not in neutrophils from *Gpr43*^{-/-} mice (Fig. 2b), indicating that GPR43 is the sole functional receptor for SCFAs on neutrophils. GPR43 has previously been reported to act as a chemoattractant receptor for acetate¹¹, and we found that neutrophils from wild-type, but not *Gpr43*^{-/-} mice, responded chemotactically to acetate, but only at very high concentrations (~100–1,000 μ M), and with a relatively low chemotactic index (Fig. 2c). This may relate to the low affinity of GPR43 for SCFAs²⁰, the high concentrations of SCFAs normally present in tissues (0.1–10 mM in serum, and 200 mM in the colon)^{11,13}, and the need for chemoattractant receptors to sense a gradient. However, a recently reported synthetic agonist specific for human GPR43 with >100-fold potency over SCFAs²⁰ was much more robust in chemotaxis assays (Fig. 2c). In neutrophils from wild-type mice, but not *Gpr43*^{-/-} mice, SCFAs induced release of ROS and increased phagocytic activity (Fig. 2d, e) similar to activities reported for certain other chemoattractant receptors.

The characteristic expression pattern of GPR43 to cell types involved in innate immunity and inflammation, and the GPR43-dependent effects of SCFA neutrophil function suggested that GPR43 may be the relevant receptor on immune cells for the regulation of inflammatory responses by SCFAs. We induced colitis by adding DSS to the drinking water of *Gpr43*^{-/-} and wild-type littermate mice. In the acute phase (7 days), *Gpr43*^{-/-} mice showed a marked increase in their inflammatory response, compared to wild-type mice (Fig. 2f), including a reduced colon length, and an increased DAI, more severe inflammation by histological analysis and increased MPO activity in the colon, indicating increased neutrophil infiltration/activation (Fig. 2f). We next determined whether acetate protected against colitis in the acute DSS model, in a GPR43-dependent manner. Mice fed 200 mM acetate in their drinking water showed a substantial decrease in inflammation, as judged by longer colon length, a reduced DAI, reduced inflammatory infiltrate and less tissue damage, when compared to wild-type mice not fed acetate (Fig. 2f). Notably, this protection occurred by acetate binding to GPR43, because acetate had no beneficial effect in *Gpr43*^{-/-} mice (Fig. 2f).

In a chronic model of DSS-induced colitis, *Gpr43*^{-/-} mice showed greater morbidity (at day 8, Fig. 2g, inset), and a marked reduction in the ability to regain weight compared to wild-type littermates (Fig. 2g). By day 42, *Gpr43*^{-/-} mice showed reduced colon length, increased colon histology score, as well as increased MPO levels in the colon. In a T-cell-dependent model of colitis, the TNBS (trinitrobenzoic sulphonic acid)-induced model, *Gpr43*^{-/-} mice also had more severe disease with decreased colon length and increased colon histological scores (Supplementary Fig. 5). CD44⁺ IL17⁺ T cells in the mesenteric lymph nodes in this model were increased in *Gpr43*^{-/-} mice, as were transcripts for IL17A, IL6, IL1 β , IFN γ and CCL2 in colon tissue (Supplementary Fig. 5).

GPR43 is also expressed on colonic epithelium, and SCFAs affect several functions of these cells including proliferation, and epithelial barrier function²¹ and butyrate is a major source of energy for colonocytes. We determined the contribution of immune and non-immune cells to the colitis phenotype we observed in *Gpr43*^{-/-} mice by using bone marrow chimaeras (Supplementary Fig. 6). After 7 weeks of bone marrow reconstitution, colitis was induced by DSS in the drinking water. Wild-type mice reconstituted with *Gpr43*^{-/-} bone marrow showed a similar exacerbated inflammatory response in the colon, compared to *Gpr43*^{-/-} mice receiving *Gpr43*^{-/-} bone marrow, demonstrating that immune cells were largely responsible for the phenotype observed in *Gpr43*^{-/-} mice.

SCFA levels in the colon can be high, particularly after ingestion of large amounts of fibre, and SCFAs are rapidly absorbed and distribute systemically through the blood²². We therefore assessed peripheral inflammatory responses in *Gpr43*^{-/-} mice using the K/BxN serum-induced model of inflammatory arthritis and an ovalbumin (OVA)-induced model of allergic airway inflammation. In the inflammatory arthritis model *Gpr43*^{-/-} mice showed a slight delay in the onset of symptoms, but by day 11 inflammation in *Gpr43*^{-/-} mice was markedly more severe than in wild-type littermates, clinically and histologically, and did not resolve over the 28 days of study (Fig. 3a). *Gpr43*^{-/-} mice showed increased levels of MPO production by peripheral blood cells (Fig. 3a). Acetate in the drinking water 1 week before, and during, the induction of inflammatory arthritis reduced inflammation (Supplementary Fig. 7a). Germ-free mice given K/BxN serum also showed increased inflammation and a much slower resolution of inflammation when compared to CNV housed mice (Supplementary Fig. 7b). *Gpr43*^{-/-} mice also showed more severe inflammation compared with

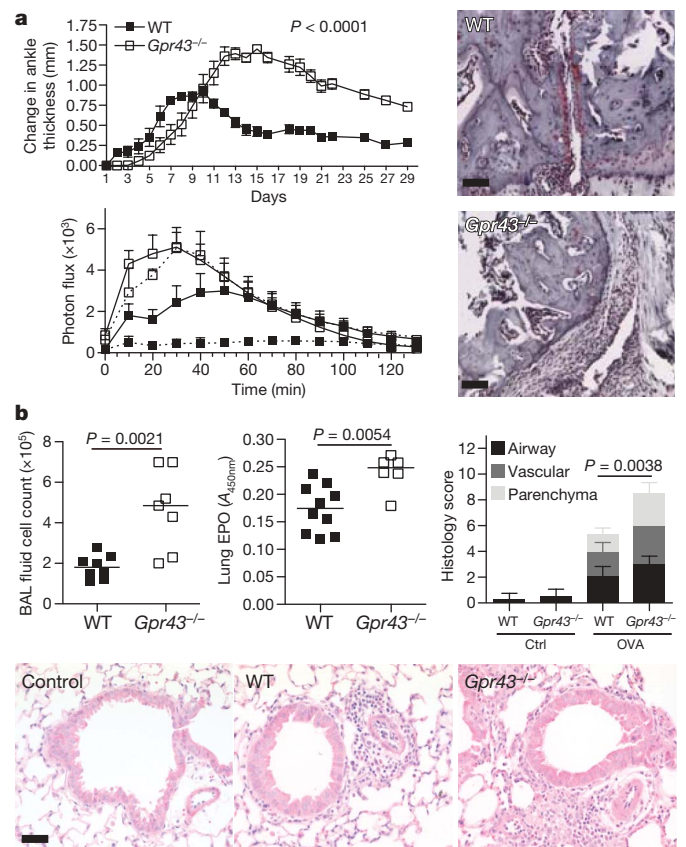


Figure 3 | Inflammatory arthritis and allergic airway disease and GPR43 deficiency. **a**, Inflammatory arthritis (K/BxN serum injection on day 0 and 2) in *Gpr43*^{-/-} mice ($n = 5$) versus wild-type littermates ($n \geq 3$). Scores shown are mean \pm s.e.m. for each time point, representative of three independent experiments. Wild-type mice are represented with closed squares, *Gpr43*^{-/-} mice with open squares, controls with dashed lines, and arthritic mice with solid lines. Change in ankle thickness (top) and measurement of MPO in the peripheral blood (bottom) showed that both naive and arthritic *Gpr43*^{-/-} mice had higher MPO production when stimulated with phorbol 12-myristate 13-acetate (PMA; bottom), indicating greater neutrophil activation ($P < 0.001$ *Gpr43*^{-/-} control compared to wild-type control, $P = 0.0019$ *Gpr43*^{-/-} compared to wild-type arthritic). Histological assessment at day 18 (right) (scale bars, 50 μ m). **b**, OVA-induced allergic airway inflammation. BAL fluid cell counts (left), eosinophil peroxidase (EPO) activity in lung tissue (middle), and inflammation as scored by histology (right). The bottom panel shows representative haematoxylin-and-eosin-stained lung sections from wild-type and *Gpr43*^{-/-} mice, and control (no OVA) mice. Scale bar, 50 μ m.

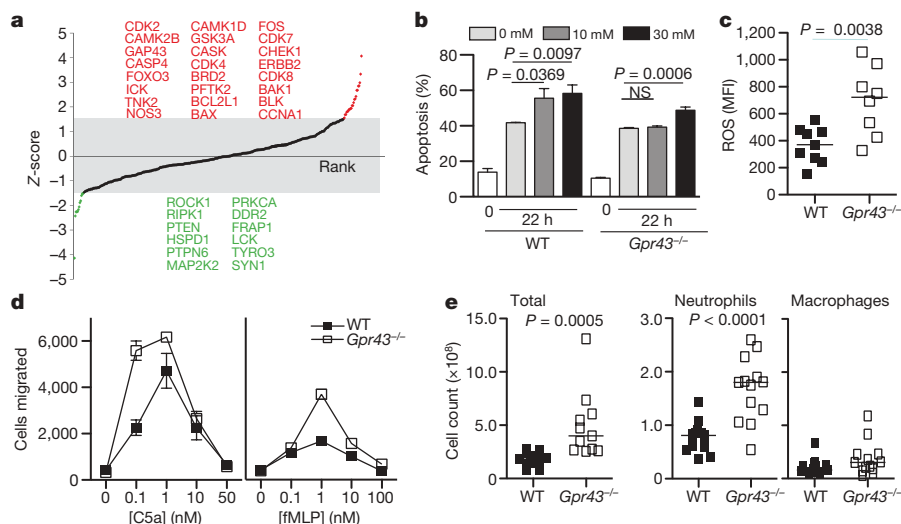


Figure 4 | GPR43 signalling and immune cell functions. **a**, Protein expression analysis using Kinex antibody microarrays. Z-score-transformed values reflecting positive or negative shifts in differential protein expression fold-changes after acetate treatment of neutrophils from wild-type mice compared to that from *Gpr43*^{-/-} mice. Proteins highlighted in red or green indicate those with Z-scores above +1.5 or below -1.5, respectively. **b**, Apoptosis in wild-type and *Gpr43*^{-/-} bone marrow cells, with or without

wild-type littermates in an acute allergic airway inflammation model, with increased numbers of cells in the broncho-alveolar lavage (BAL) fluid, as well as greater levels of eosinophil peroxidase and inflammatory cells in the lung tissue (Fig. 3b).

The cellular and molecular basis for GPR43–SCFA effects on inflammatory responses was studied using Kinex protein microarrays that interrogate more than 600 proteins and phosphoproteins. Because SCFAs also inhibit histone deacetylases and thereby affect cell transcription and functions²³, we made direct comparisons between wild-type and *Gpr43*^{-/-} neutrophils, to identify GPR43-related signalling pathways affected by SCFAs. Protein–protein interaction networks were constructed from the set of top-ranked proteins exhibiting greatest positive or negative differential shifts (Z-scores ≥ 1.5 and ≤ -1.5 ; Fig. 4a). We applied the MCODE clustering algorithm to identify highly connected local sub-networks, which revealed interesting modules such as the apoptosis-associated BAX–BAK1–BCL2L1 cluster and the PRKCA–PTPN6–LCK cluster (Supplementary Fig. 8a). Consistent with the changes in apoptosis-related signalling molecules, acetate induced apoptosis in neutrophils in a dose-dependent and a GPR43-dependent manner (Fig. 4b), except at very high concentrations (30 mM). There were also differences between wild-type and *Gpr43*^{-/-} neutrophils with respect to several other signalling pathways associated with inflammation, cell migration or apoptosis in mouse, and in human (Supplementary Figs 9 and 10). Granulocytes from *Gpr43*^{-/-} mouse blood showed an increase in levels of ROS (Fig. 4c) and MPO (Fig. 3a). Interestingly, macrophages from germ-free mice also show increased production of ROS²⁴. *Gpr43*^{-/-} neutrophils also showed increased chemotaxis to the *N*-formyl peptide f-Met-Leu-Phe (fMLP) and to the complement fragment C5a (Fig. 4d), compared to wild-type cells. Furthermore, heat-inactivated *Staphylococcus aureus* injected into the peritoneum of *Gpr43*^{-/-} mice yielded a greater recruitment of neutrophils after 1 h, compared to wild-type mice (Fig. 4e). Acetate stimulation of human neutrophils markedly reduced surface expression of pro-inflammatory receptors such as C5aR and CXCR2 (Supplementary Fig. 11), presumably through agonist-mediated receptor heterodimerization and internalisation.

Commensal bacteria and vertebrate immune systems form a symbiotic relationship and have co-evolved²⁵ such that proper immune development and function relies on colonisation of the gastrointestinal tract by commensal bacteria^{2,26}. SCFA–GPR43 signalling

acetate stimulation (apoptotic cells are annexin V and propidium iodide (PI) double positive). **c**, Chemotactic response to fMLP and C5a by wild-type and *Gpr43*^{-/-} bone marrow granulocytes. **d**, Recruitment of neutrophils and macrophages to the peritoneum in wild-type and *Gpr43*^{-/-} mice, injected with 1×10^6 heat-inactivated *S. aureus* particles. **e**, Reactive oxygen species production by peripheral blood granulocytes.

is one of the molecular pathways whereby commensal bacteria regulate immune and inflammatory responses. GPR43 resembles another anti-inflammatory chemoattractant receptor, ChemR23 (ref. 27), although this receptor binds endogenous rather than bacterially produced ligands. Any agents that affect gastrointestinal microbiota, and the production of SCFAs, might be expected to influence immune and inflammatory responses. It is possible that high levels of SCFAs such as acetate may, in addition to its direct effects on the GPR43 response, affect the biosynthesis of endogenous fatty acids, such as resolvins, that modulate leukocyte functions. Levels of SCFAs in the gastrointestinal tract vary significantly depending on the amount of non-digestible fibre in the diet, and also relate to the composition of the gut microbiota¹². For instance, the relative proportion of Bacteroidetes is decreased in obese people compared to lean people²⁸. Indeed an altered composition of the gut microbiota, brought on by western diet, or by use of antibiotics, has been suggested as a reason for the increased incidence of allergies and asthma in humans²⁹. SCFA–GPR43 interactions could represent a central mechanism to account for affects of diet, prebiotics and probiotics on immune responses, and may represent new avenues for understanding and potentially manipulating immune responses.

METHODS SUMMARY

RNA extraction, preparation, hybridization and expression analysis using U133A and B Affymetrix GeneChips (Gene Expression Omnibus (GEO) accession GSE3982) were performed as previously described, using an extensive collection of immune cell data sets¹⁸. Protein expression and phosphorylation were assessed using Kinex antibody microarrays (Kinex Bioinformatics) (<http://www.kinex.ca/services>). *Gpr43*^{-/-} mice on a C57BL/6 background were obtained from Deltagen (<http://www.deltagen.com>). The K/BxN inflammatory arthritis model, the DSS and TNBS models of colitis, and the OVA allergic airway inflammation followed standard published procedures. Statistical analyses were conducted using a Student's two-way *t*-test, or two-way analysis of variance (ANOVA) using Graphpad Prism software.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 7 August; accepted 18 September 2009.

- Wen, L. *et al.* Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* **455**, 1109–1113 (2008).

2. Mazmanian, S. K., Round, J. L. & Kasper, D. L. A microbial symbiosis factor prevents intestinal inflammatory disease. *Nature* **453**, 620–625 (2008).
3. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl Acad. Sci. USA* **104**, 13780–13785 (2007).
4. Treem, W. R., Ahsan, N., Shoup, M. & Hyams, J. S. Fecal short-chain fatty acids in children with inflammatory bowel disease. *J. Pediatr. Gastroenterol. Nutr.* **18**, 159–164 (1994).
5. Harig, J. M., Soergel, K. H., Komorowski, R. A. & Wood, C. M. Treatment of diversion colitis with short-chain-fatty acid irrigation. *N. Engl. J. Med.* **320**, 23–28 (1989).
6. Kanauchi, O. *et al.* Treatment of ulcerative colitis by feeding with germinated barley foodstuff: first report of a multicenter open control trial. *J. Gastroenterol.* **37** (suppl. 14), 67–72 (2002).
7. Breuer, R. I. *et al.* Rectal irrigation with short-chain fatty acids for distal ulcerative colitis. Preliminary report. *Dig. Dis. Sci.* **36**, 185–187 (1991).
8. Scheppach, W. Treatment of distal ulcerative colitis with short-chain fatty acid enemas. A placebo-controlled trial. German-Austrian SCFA Study Group. *Dig. Dis. Sci.* **41**, 2254–2259 (1996).
9. Vernia, P. *et al.* Short-chain fatty acid topical treatment in distal ulcerative colitis. *Aliment. Pharmacol. Ther.* **9**, 309–313 (1995).
10. Brown, A. J. *et al.* The Orphan G protein-coupled receptors GPR41 and GPR43 are activated by propionate and other short chain carboxylic acids. *J. Biol. Chem.* **278**, 11312–11319 (2003).
11. Le Poul, E. *et al.* Functional characterization of human receptors for short chain fatty acids and their role in polymorphonuclear cell activation. *J. Biol. Chem.* **278**, 25481–25489 (2003).
12. Macfarlane, S. & Macfarlane, G. T. Regulation of short-chain fatty acid production. *Proc. Nutr. Soc.* **62**, 67–72 (2003).
13. Høverstad, T. & Midtvedt, T. Short-chain fatty acids in germfree mice and rats. *J. Nutr.* **116**, 1772–1776 (1986).
14. Tedelind, S., Westberg, F., Kjerrulf, M. & Vidal, A. Anti-inflammatory properties of the short-chain fatty acids acetate and propionate: a study with relevance to inflammatory bowel disease. *World J. Gastroenterol.* **13**, 2826–2832 (2007).
15. Cavaglieri, C. R. *et al.* Differential effects of short-chain fatty acids on proliferation and production of pro- and anti-inflammatory cytokines by cultured lymphocytes. *Life Sci.* **73**, 1683–1690 (2003).
16. Segain, J. P. *et al.* Butyrate inhibits inflammatory responses through NF- κ B inhibition: implications for Crohn's disease. *Gut* **47**, 397–403 (2000).
17. Lührs, H. *et al.* Butyrate-enhanced TNF α -induced apoptosis is associated with inhibition of NF- κ B. *Anticancer Res.* **22**, 1561–1568 (2002).
18. Jeffrey, K. L. *et al.* Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. *Nature Immunol.* **7**, 274–283 (2006).
19. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
20. Lee, T. *et al.* Identification and functional characterization of allosteric agonists for the G protein-coupled receptor FFA2. *Mol. Pharmacol.* **74**, 1599–1609 (2008).
21. Suzuki, T., Yoshida, S. & Hara, H. Physiological concentrations of short-chain fatty acids immediately suppress colonic epithelial permeability. *Br. J. Nutr.* **100**, 297–305 (2008).
22. Pomare, E. W., Branch, W. J. & Cummings, J. H. Carbohydrate fermentation in the human colon and its relation to acetate concentrations in venous blood. *J. Clin. Invest.* **75**, 1448–1454 (1985).
23. Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature* **389**, 349–352 (1997).
24. Mørland, B. & Midtvedt, T. Phagocytosis, peritoneal influx, and enzyme activities in peritoneal macrophages from germfree, conventional, and ex-germfree mice. *Infect. Immun.* **44**, 750–752 (1984).
25. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).
26. Rakoff-Nahoum, S., Paglino, J., Eslami-Varzaneh, F., Edberg, S. & Medzhitov, R. Recognition of commensal microflora by toll-like receptors is required for intestinal homeostasis. *Cell* **118**, 229–241 (2004).
27. Serhan, C. N., Chiang, N. & Van Dyke, T. E. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nature Rev. Immunol.* **8**, 349–361 (2008).
28. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023 (2006).
29. Shreiner, A., Huffnagle, G. B. & Nover, M. C. The “Microflora Hypothesis” of allergic disease. *Adv. Exp. Med. Biol.* **635**, 113–134 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank P. Silvera and S. Tangye for supply of certain Genechip data sets, L. Tsai for help with heatmaps, and D. Kobuley, J. Nicoli and M. Abt for help in the germ-free animal facilities. K.M.M. and C.R.M. are supported by the Australian NHMRC, and the CRC for Asthma and Airways. A.N. is a recipient of a Fellowship award from the Crohn's and Colitis Foundation of America. F.S. and D.Y. are Cancer Institute NSW Fellows.

Author Contributions C.R.M. conceived and supervised the project, and K.M.M. performed the vast majority of the *in vitro* and *in vivo* experiments (other than those detailed below) and provided intellectual input to scientific direction and interpretations. A.T.V., M.M.T. and D.A. contributed to experiments with germ-free mice. F.M., M.S.R. and F.S. identified GPR43 as a receptor with an interesting transcript expression, and A.N. and R.J.X. were responsible for all of the subsequent bioinformatic analyses. H.C.S., D.Y. and J.K. provided general support for many of the experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.R.M. (charles.mackay@med.monash.edu.au).

METHODS

Microarray gene expression analysis. RNA extraction, preparation, hybridization and expression analysis using U133A and B Affymetrix GeneChips were performed as previously described, using an extensive immune cell data set³⁰. For analysis of immune cell data sets, MAS5-normalized data was filtered to remove probe sets identified as 'absent' (MAS5 algorithm) across all samples. The data was further filtered by setting a minimum threshold value >20 in at least one sample for each probe set and a maximum-minimum expression value >100. The data was then log- and Z-score-transformed. Pearson correlation coefficients were calculated for each pair of genes as well as for *GPR43*, and each gene represented on the array and k-means clustering performed³¹. Heatmaps were generated using TreeView³².

Kinex protein microarrays. Preparation of samples was done according to standard Kinex recommendations (<http://www.kinexus.ca/services>). The normalized data from Kinex was further filtered to remove data points with error ranges exceeding Kinex's suggested threshold of 15% for adjacent duplicate spots. Positive or negative shifts in log expression fold-changes after SCFA treatment of *Gpr43*^{-/-} cell samples compared to that from wild-type samples were Z-score-transformed, identifying only top-scoring proteins with Z-scores above +1.5 or below -1.5.

Human protein interaction networks. Protein-protein interaction networks were constructed by iteratively connecting interacting proteins using curated data from the Human Protein Reference Database (HPRD)³³. The network uses graph theoretic representations which use abstract components (gene products) as nodes and relationships (interactions) between components as edges, implemented in the Perl programming language.

Neutrophil assays. Human neutrophils were isolated from the peripheral venous blood of healthy volunteers, using 1% dextran sedimentation for 30 min followed by centrifugation over 65% Percoll (Amersham Bioscience). Mouse neutrophils were isolated from hind leg femurs and separated by density centrifugation over Ficoll-Paque (Amersham Bioscience).

Reactive oxygen species were detected using H₂DCFDA (Sigma), as per manufacturer's directions. Twenty-five microlitres of 10 μ M DCFDA was incubated with 25 μ l whole blood for 20 min at 37 °C. Cells were then fixed, and red blood cells lysed. Cells were resuspended in FACS buffer.

Phagocytosis was assessed using fluorescently labelled Bioparticles, *Escherichia coli* K-12 strain or *S. aureus* (wood strain without protein A)-BODIPYFL conjugates (Molecular Probes). *In vitro*: 25 μ l of whole blood was incubated with 3 \times 10⁵ Bioparticles for 30 min, then analysed by flow cytometry. Chemotaxis was assessed using 3 μ M Multiscreen-MIC 96-well plates (Millipore). Bone marrow neutrophils (3–5 \times 10⁵) were in the top chamber, with chemoattractants in bottom chamber (fMLP, C5a, acetate, concentrations as shown in figures), cells and chemoattractants were in Chemotaxis buffer (50:50 RPMI and M199 + 2% BCS). Cell migration was assessed by flow cytometry as described previously³⁴. Apoptosis was assessed using Annexin V-FITC and propidium iodide as per manufacturer's instructions (Becton Dickinson).

Animals and models. All experimental procedures involving mice were carried out according to protocols approved by the relevant Animal Ethics Committees. *Gpr43*^{-/-} mice on a C57Bl/6 background were obtained from Deltagen. Housing and maintenance of germ-free mice followed previously described procedures^{35,36}. Germ-free mice were raised and housed germ free, as described previously³⁷.

Acute DSS colitis: mice were fed DSS in drinking water (percentage indicated in each figure) for 7 days and were monitored daily for stool score and visible bleeding. Stools were scored between 0 and 4, with 0 being normal up to 4 being diarrhoea, as described previously³⁸. At 7 days mice were killed and colons measured and assessed histologically using standard procedures. For bone marrow chimeras, *Gpr43*^{-/-} (CD45.2) and congenic C57Bl/6 wild-type mice (CD45.1) were irradiated with 9 Gy then reconstituted with donor bone marrow (10⁷ cells) for 7 weeks. Reconstitution was checked by flow cytometry, using CD45 to determine bone marrow origin. For acetate-treatment experiments mice were fed acetate in the drinking water (concentration indicated in figure legends) for 5–7 days before DSS administration, and also in combination with DSS. MPO was detected as previously described³⁹. Colon

histology scoring was performed from adapted methods previously described. Damage was scored using the following system. Mucosal ulceration: 0, no injury; 1, focal injury; 2, multifocal injury; 3, diffuse ulceration/infiltration. Depth of injury: 0, no injury; 1, mucosal involvement only; 2, mucosal and submucosal involvement; 3, transmural involvement.

Chronic DSS colitis: Mice were fed DSS on days 0–6 (4%), 20–24 (2%) and 35–39 (4%), and killed on day 42. Mice were given normal drinking water in the rest periods. Disease was monitored as per the acute model.

TNBS-induced colitis: mice were sensitised by applying a mixture of acetone/olive oil (50:50) with TNBS (Sigma) (50:50, total) on shaved skin between shoulder blades. Seven days later, mice were challenged intra-rectally with 2.5 mg TNBS with 50% ethanol, 3.5 cm from the anal verge. Mice were fasted overnight before the intra-rectal challenge, and given 5% dextrose in the drinking water. Mice were killed 3 days after TNBS challenge.

For the K/BxN inflammatory arthritis model, 150 μ l serum from K/BxN arthritic mice was used to induce experimental arthritis in recipient mice, and disease progression was monitored^{40,41}. The clinical score was calculated for each mouse by summing the scores for the four paws: 0, normal joint; 1, mild-to-moderate swelling of the ankle and/or one swollen digit; 2, swollen ankle or swelling in two or more digits; 3, severe swelling along all aspects of paw or all five digits swollen. For SCFA-feeding, sodium acetate was dissolved in drinking water at 200 mM (or as indicated) and fed to mice 1 week before colitis or arthritis induction, as well as during disease monitoring. Myeloperoxidase bioluminescence was detected in peripheral blood as described previously⁴².

Allergic airway disease was induced using OVA/alum as previously described⁴³ with modification. In brief, OVA/alum was injected interperitoneally on days 0 and 7, followed by intranasal challenge on days 12 to 15 and mice were killed on day 16. Eosinophil peroxidase was detected as previously described⁴⁴.

30. Liu, S. M. *et al.* Immune cell transcriptome datasets reveal novel leukocyte subset-specific genes and genes associated with allergic processes. *J. Allergy Clin. Immunol.* **118**, 496–503 (2006).
31. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
32. Saldanha, A. J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
33. Mishra, G. R. *et al.* Human protein reference database—2006 update. *Nucleic Acids Res.* **34**, D411–D414 (2006).
34. Heath, H. *et al.* Chemokine receptor usage by human eosinophils. The importance of CCR3 demonstrated using an antagonistic monoclonal antibody. *J. Clin. Invest.* **99**, 178–184 (1997).
35. Souza, D. G. *et al.* The essential role of the intestinal microbiota in facilitating acute inflammatory responses. *J. Immunol.* **173**, 4137–4146 (2004).
36. Zaph, C. *et al.* Commensal-dependent expression of IL-25 regulates the IL-23–IL-17 axis in the intestine. *J. Exp. Med.* **205**, 2191–2198 (2008).
37. Souza, D. G. *et al.* The required role of endogenously produced lipoxin A4 and annexin-1 for the production of IL-10 and inflammatory hyporesponsiveness in mice. *J. Immunol.* **179**, 8533–8543 (2007).
38. Cooper, H. S., Murthy, S. N., Shah, R. S. & Sedergran, D. J. Clinicopathologic study of dextran sulfate sodium experimental murine colitis. *Lab. Invest.* **69**, 238–249 (1993).
39. Vieira, A. T. *et al.* Mechanisms of the anti-inflammatory effects of the natural secosteroids physalins in a model of intestinal ischaemia and reperfusion injury. *Br. J. Pharmacol.* **146**, 244–251 (2005).
40. Korganow, A. S. *et al.* From systemic T cell self-reactivity to organ-specific autoimmune disease via immunoglobulins. *Immunity* **10**, 451–461 (1999).
41. Lee, D. M. *et al.* Mast cells: a cellular link between autoantibodies and inflammatory arthritis. *Science* **297**, 1689–1692 (2002).
42. Gross, S. *et al.* Bioluminescence imaging of myeloperoxidase activity *in vivo*. *Nature Med.* **15**, 455–461 (2009).
43. Shum, B. O. *et al.* The adipocyte fatty acid-binding protein aP2 is required in allergic airway inflammation. *J. Clin. Invest.* **116**, 2183–2192 (2006).
44. Strath, M., Warren, D. J. & Sanderson, C. J. Detection of eosinophils using an eosinophil peroxidase assay. Its use as an assay for eosinophil differentiation factors. *J. Immunol. Methods* **83**, 209–215 (1985).

Resolvin D2 is a potent regulator of leukocytes and controls microbial sepsis

Matthew Spite¹, Lucy V. Norling^{1,2}, Lisa Summers¹, Rong Yang¹, Dianne Cooper², Nicos A. Petasis³, Roderick J. Flower², Mauro Perretti² & Charles N. Serhan¹

A growing body of evidence indicates that resolution of acute inflammation is an active process^{1,2}. Resolvins are a new family of lipid mediators enzymatically generated within resolution networks that possess unique and specific functions to orchestrate catabasis, the phase in which disease declines^{2,3}. Resolvin D2 (RvD2) was originally identified in resolving exudates, yet its individual contribution in resolution remained to be elucidated. Here, we establish RvD2's potent stereoselective actions in reducing excessive neutrophil trafficking to inflammatory loci. RvD2 decreased leukocyte–endothelial interactions *in vivo* by endothelial-dependent nitric oxide production, and by direct modulation of leukocyte adhesion receptor expression. In mice with microbial sepsis initiated by caecal ligation and puncture, RvD2 sharply decreased both local and systemic bacterial burden, excessive cytokine production and neutrophil recruitment, while increasing peritoneal mononuclear cells and macrophage phagocytosis. These multi-level pro-resolving actions of RvD2 translate to increased survival from sepsis induced by caecal ligation and puncture and surgery. Together, these results identify RvD2 as a potent endogenous regulator of excessive inflammatory responses that acts via multiple cellular targets to stimulate resolution and preserve immune vigilance.

Ungoverned inflammation is an underlying component of many pathologies, such as cardiovascular disease, diabetes and sepsis^{4,5}. It's now recognized that resolution of inflammation is an active programme controlled by temporal and spatial production of specialized chemical mediators^{2,3,6}. Recently, autacoids endogenously generated from omega-3 essential fatty acids, namely resolvins, were identified during the resolution phase of inflammation that actively promote catabasis via potent pro-resolving and anti-inflammatory actions^{2,3}. RvD2, biosynthesized from docosahexaenoic acid (DHA), was originally identified during resolution³. Its complete stereochemistry and actions remained of interest. To this end, we investigated whether RvD2 preserves host immune function to facilitate resolution of inflammatory sepsis.

First, the complete stereochemistry of endogenous RvD2 was determined by physical matching with compounds prepared by total organic synthesis (Supplementary Fig. 1a) from enantiomerically and geometrically pure starting materials in accordance with the basic structure determined in resolving exudates³ (Fig. 1). This approach was needed because the nanogram amounts of endogenous RvD2 isolated precluded direct nuclear magnetic resonance (NMR) analysis. The double-bond geometry of synthetic material was validated by ¹H NMR (Supplementary Fig. 1b). The biosynthesis of RvD2 involves 17-lipoxygenation of DHA to 17S-hydroperoxy-4Z, 7Z, 10Z, 13Z, 15E, 19Z-docosahexaenoic acid (17-HpDHA), which is enzymatically transformed to a 7(8)epoxide-containing intermediate^{3,7} in human

leukocytes. This enzymatic activity involves 5-lipoxygenase (LOX) and its epoxide-generating activity⁸. These steps can occur within a single cell type or via transcellular biosynthesis. For example, eosinophils, rich in 15-LOX, can convert DHA to 17-HpDHA, which polymorphonuclear neutrophils (PMNs) can convert to RvD2 (Fig. 1a). Actively phagocytosing PMNs converted resolvin precursor 17-HpDHA to RvD2 as determined by lipidomics based on liquid-chromatography tandem mass spectrometry (LC-MS/MS). A total ion chromatogram (mass-to-charge ratio $m/z = 375$ [M-H]) of human leukocyte-derived RvD2 is shown (Fig. 1b), with characteristic conjugated tetraene ultraviolet (UV)-chromophore (absorbance λ_{\max} at 301 nm with shoulders at 289 nm and 315 nm). Synthetic material showed an exclusive and prominent peak with retention time and UV spectrum essentially identical to leukocyte-derived RvD2 (Fig. 1c). Co-injection of synthetic and leukocyte-derived RvD2 led to an increase in intensity and co-elution (Fig. 1d). To further establish the physical properties, their tandem mass spectra were analysed, with essentially identical mass spectrum, and diagnostic ions in agreement with original assignments for endogenous RvD2 (Fig. 1e, f)³. To validate the biosynthetic pathway, activated human PMNs were incubated with deuterium-labelled 17S-HpDHA-d₅ or DHA-d₅; RvD2 containing the d₅ label was biosynthesized (Fig. 1g), the parent ion [M-H] increased to m/z 380, and neutral loss ions reflective of d₅-containing fragments (Supplementary Fig. 2). Next, to further confirm the structural assignment^{3,9}, derivatized RvD2 was subjected to gas chromatography-mass spectrometry (GC/MS). The derivatized RvD2 C-value was 25.2 ± 0.1 and its spectrum (Supplementary Fig. 3) showed diagnostic ions at m/z 479, 435, 229 and 171 (see ref. 3). Collectively, matching of synthetic and leukocyte-derived RvD2 (by retention time, mass spectral diagnostic ions and derivatization) established the complete stereochemistry and double-bond geometry of endogenous RvD2 as 7S, 16R, 17S-trihydroxy-4Z, 8E, 10Z, 12E, 14E, 19Z-docosahexaenoic acid.

RvD2 displayed potent actions in microbial peritonitis, with a drastic ~70% reduction in zymosan-stimulated PMN infiltration at doses as low as 10 pg (Fig. 2). Importantly, the $\Delta 10$ -*trans*-RvD2 isomer was essentially inactive, indicating that specific geometry of endogenous RvD2 is required for bioactivity. To determine whether RvD2 decreases leukocyte–endothelial interactions, microcirculation in the cremaster muscle was analysed. Platelet activating factor (PAF; 100 nM)^{6,10} superfusion caused increased leukocyte adherence and emigration that was markedly reduced by 1 nM RvD2 (Fig. 2c, d). Representative microcirculation before and after RvD2 superfusion is shown in the Supplementary movies. These potent RvD2 actions were recapitulated with human cells to identify potential cellular directed actions (that is, endothelial cells and PMN). RvD2 potently reduced PAF-stimulated

¹Center for Experimental Therapeutics and Reperfusion Injury, Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. ²William Harvey Research Institute, Barts and the London Medical School, Queen Mary University of London, London EC1M 6BQ, UK. ³Department of Chemistry and Loker Hydrocarbon Research Institute, University of Southern California, Los Angeles, California 90089, USA.

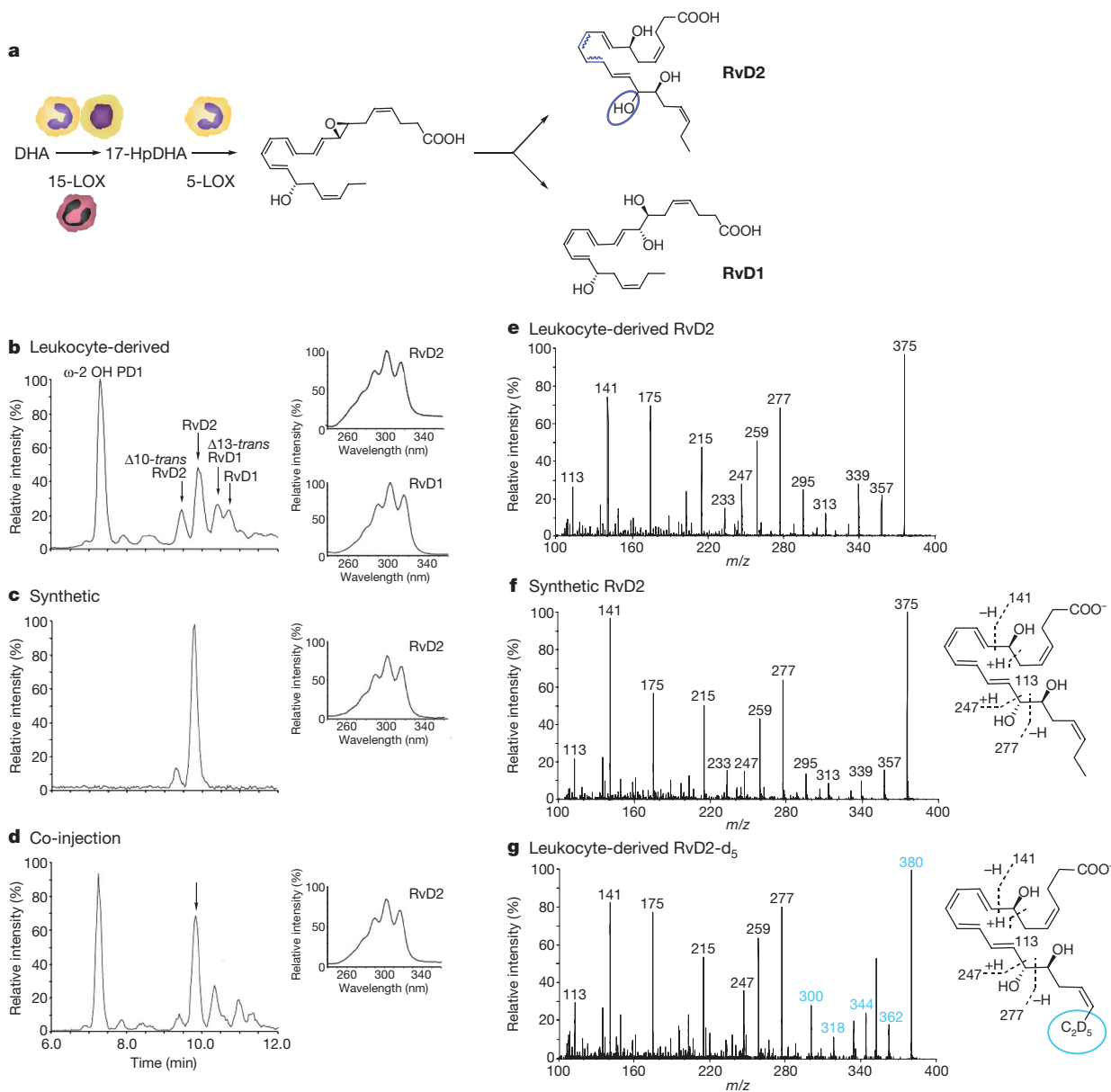


Figure 1 | Stereochemical assignment, biosynthesis and total organic synthesis of RvD2. **a**, Rv biosynthesis illustrating potential PMN and eosinophil transcellular biosynthesis. Blue, stereochemistries for assignments. **b**, Ion chromatograms (m/z 375) depicting human leukocyte-derived RvD2 and RvD1 with UV spectra (insets) and related isomers with the ω -2 OH metabolites of PD1 and 7,17-diHDHA. **c**, **d**, RvD2 prepared by

total organic synthesis (**c**) and co-injection with leukocyte-derived RvD2 (**d**). **e**, **f**, MS/MS of leukocyte-derived RvD2 (**e**) and synthetic RvD2 (**f**) with prominent ions at m/z 357 [M-H-H₂O], 339 [M-H-2H₂O], 313 [M-H-CO₂-H₂O], 295 [M-H-CO₂-2H₂O], 277, 259 [277-H₂O], 247, 233 [277-CO₂], 141 and 113. **g**, Leukocyte-derived RvD2- d_5 . Blue, d_5 -labelled ions. Representative of $n = 3-5$.

capture and adhesion of PMN by human umbilical vein endothelial cells (HUVECs) under flow (Supplementary Fig. 4)¹¹. We note that RvD2 also reduced complement-mediated (with complement C5a) PMN-endothelial interactions (Supplementary Fig. 5a-c), a key mediator in sepsis¹². Consistent with the impact of RvD2 on leukocyte-endothelial interactions, RvD2 diminished PAF-stimulated CD62L (L-selectin) shedding on isolated human PMN, and CD18 (ITGB2, also known as integrin beta 2) surface expression (Fig. 2e, f). RvD2 alone did not alter PMN adhesion molecule expression ($n = 3$, not shown). To obtain further evidence for direct actions of RvD2 on human leukocytes, we monitored reactive oxygen species (ROS). Importantly, RvD2 did not stimulate extracellular superoxide, and it potently reduced C5a-stimulated extracellular superoxide generation (see below and Supplementary Fig. 5e, f).

Next, we assessed the contribution of nitric oxide, an established anti-adhesive mediator^{13,14} in RvD2-reduced leukocyte adhesion in post-capillary venules. The non-selective nitric oxide synthase inhibitor

L-NAME, before addition of RvD2, partially reversed the decreased leukocyte adherence and emigration (Fig. 3a, b). To obtain additional evidence for nitric oxide generation by RvD2 *in vivo*, vascular fluorescence was monitored (see Methods). Topical administration of RvD2 (100 pg per ear) increased fluorescence intensity, whereas lower doses (1 pg and 10 pg) were ineffective (Supplementary Fig. 6a). L-NAME given before topical RvD2 application abolished this response (Supplementary Fig. 6b), indicating that RvD2-stimulated vascular responses at this dose were nitric-oxide-dependent. We note that intravenous injection of RvD2 at doses that inhibited PMN infiltration in peritonitis (10 pg) did not increase fluorescence intensity, indicating that only local elevated doses of RvD2 stimulated vascular responses (not shown). Additionally, RvD2 superfusion (1 nM) did not cause an increase in vascular permeability (Supplementary Fig. 6e). Topical RvD2 (10 pg or 100 pg) did not induce leukocyte infiltration into ear skin compared to chemoattractant leukotriene B₄ (Supplementary Fig. 6f). These results demonstrate that high focal delivery of RvD2 stimulates rapid

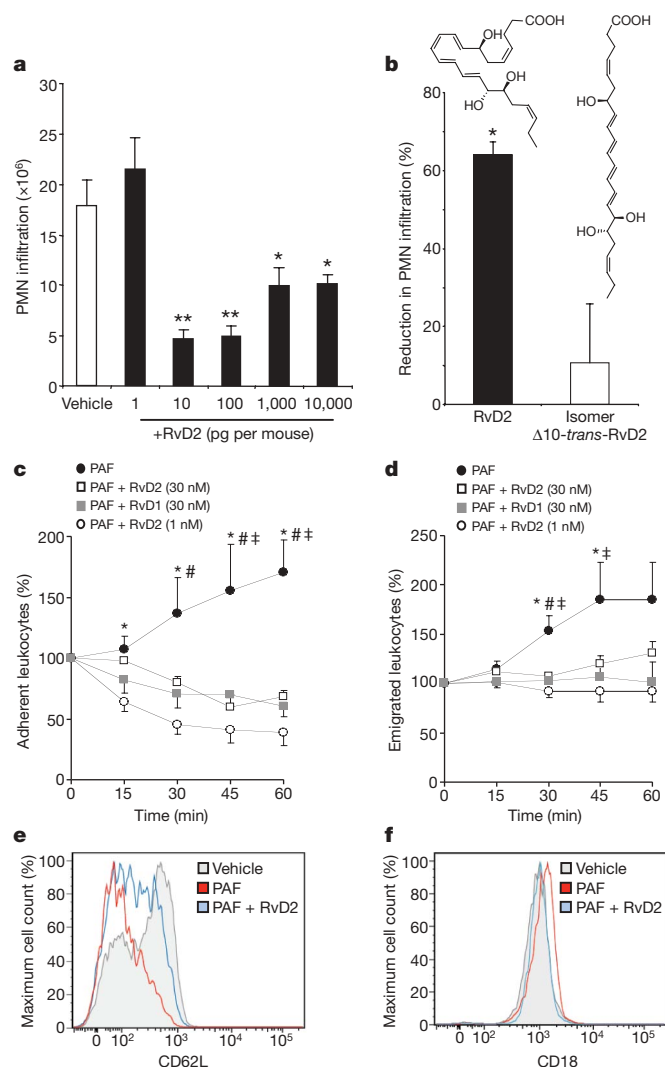


Figure 2 | RvD2 potentially reduces leukocyte-endothelial interactions to reduce microbial peritonitis. **a**, **b**, Leukocyte infiltration in peritonitis. $*P < 0.05$, $**P < 0.01$ ANOVA. **b**, Equidose comparison (100 pg) of RvD2 and $\Delta 10$ -trans-RvD2. **c**, **d**, Leukocyte trafficking *in vivo*. PAF-stimulated (100 nM) leukocyte adherence (**c**) and emigration (**d**) with or without RvD2 or RvD1. $*P < 0.01$ (PAF versus PAF + RvD2 1 nM), $\#P < 0.05$ (PAF versus PAF + RvD1 30 nM) and $\ddagger P < 0.05$ (PAF versus PAF + RvD2 30 nM) ANOVA. **e**, **f**, Adhesion receptor surface expression. Results ($n = 3-6$) are mean \pm s.e.m.

nitric oxide production consistent with its anti-adhesive effects but not to a level that is pro-inflammatory.

Corroboratory results were obtained with HUVECs whereby RvD2 dose-dependently stimulated nitric oxide generation (Fig. 3c), suggesting that topical actions were probably mediated via endothelial nitric oxide synthase (eNOS). To test this, peritonitis was evaluated in mice deficient in eNOS. In agreement with an earlier report¹⁵, no changes were observed with respect to leukocyte infiltration between wild-type and *eNOS*^{-/-} (also known as *Nos3*^{-/-}) mice. RvD2 reduction in leukocytes was eliminated in *eNOS*^{-/-} mice (Fig. 3d), an effect that has also been reported for aspirin and local aspirin-triggered lipoxins¹⁶. Notably, RvD2 also stimulated vasoprotective prostacyclin (6-keto-PGF_{1 α} ; Supplementary Fig. 7a); this dose-response curve proved to be bell-shaped like other lipid mediators^{1,2,6}. RvD2-stimulated prostacyclin and nitric oxide were sensitive to pertussis toxin, implicating a role for G-protein coupled receptor(s) (Supplementary Fig. 7b, c). Thus, RvD2 regulates leukocyte adherence via both direct actions on PMN (see above) and endothelial vasoactive substances.

Next, anti-inflammatory and pro-resolving actions of RvD2 were evaluated in caecal ligation and puncture (CLP), an established murine

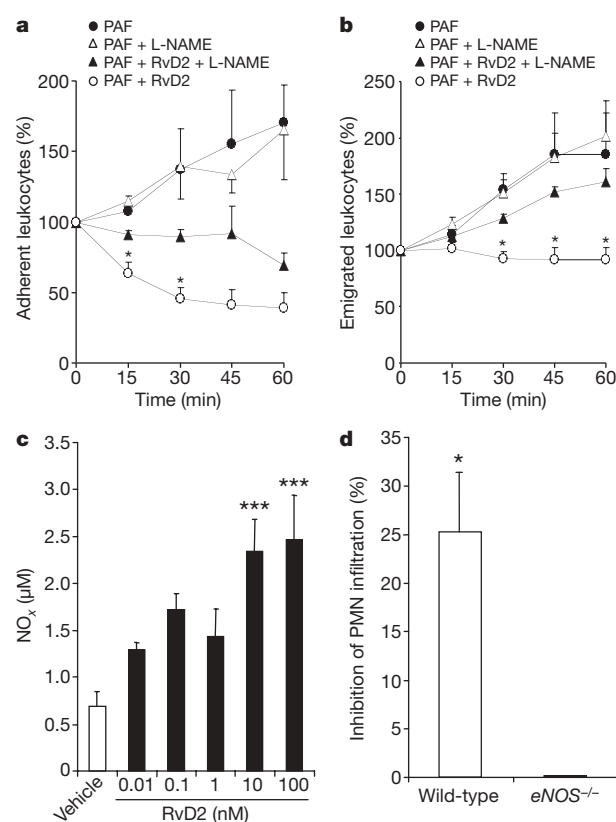


Figure 3 | Modulation of leukocyte trafficking by RvD2 is nitric oxide dependent. **a**, **b**, Leukocyte trafficking *in vivo*. Cremasters were superfused with PAF (100 nM), L-NAME (100 μ M) and RvD2 (1 nM) and leukocyte adhesion (**a**) and emigration (**b**) was quantified ($n = 3-5$). $*P < 0.05$ two-way ANOVA. **c**, Nitric oxide (NO_x; nitrate or nitrite) generation in primary HUVECs incubated with RvD2 ($n = 4-6$). $***P < 0.001$, one-way ANOVA. **d**, Zymosan-stimulated PMN infiltration after administration of RvD2 (100 ng; intravenously) in wild-type and *eNOS*^{-/-} mice ($n = 3-5$). Results are mean \pm s.e.m. $*P < 0.05$, two-tailed unpaired Student's *t*-test.

microbial sepsis that closely resembles human pathology^{17,18}. Omega-3 fatty acids are beneficial in some inflammatory conditions, including sepsis¹⁹⁻²¹, although the mechanistic basis underlying protection is still emerging. Indeed, RvD2 significantly reduced the amount of live aerobic bacteria in both blood and peritoneum at 12 h post-CLP, whereas $\Delta 10$ -trans-RvD2 was essentially not active (Fig. 4a, b). This was associated with a significant reduction in total leukocytes and specifically, PMN infiltration into the peritoneum (Fig. 4c, d). Interestingly, the ratio of mononuclear cells to PMN was increased with RvD2 (Fig. 4d, inset), similar to results obtained with sterile zymosan-stimulated peritonitis (Supplementary Fig. 9a). Intra-peritoneal delivery of RvD2 at 1 h post-CLP also reduced both blood and peritoneal bacteria (Supplementary Fig. 8). Macrophages have an important role in the clearance of bacteria, cellular debris and apoptotic PMNs to facilitate inflammation resolution^{1,2}. RvD2 treatment promoted phagocyte-dependent bacterial clearance observed in inguinal lymph nodes (Supplementary Fig. 10d). Evidence for direct macrophage actions were obtained *in vitro*, where RvD2 potentially enhanced macrophage phagocytosis of opsonized-zymosan (Supplementary Fig. 9b).

To obtain additional evidence for this pro-resolution role of RvD2, cytokines were monitored during CLP both locally (peritoneum) and systemically (plasma). RvD2 drastically reduced levels of pro-inflammatory cytokines associated with poor outcomes in sepsis¹⁸, namely the interleukins IL-6, IL-1 β , IL-23 and tumour necrosis factor TNF- α (Fig. 4e and Supplementary Fig. 10a). RvD2 reduced cytokine levels while enhancing bacterial clearance, a response also observed for macrophage scavenger receptor A²². Hence, it

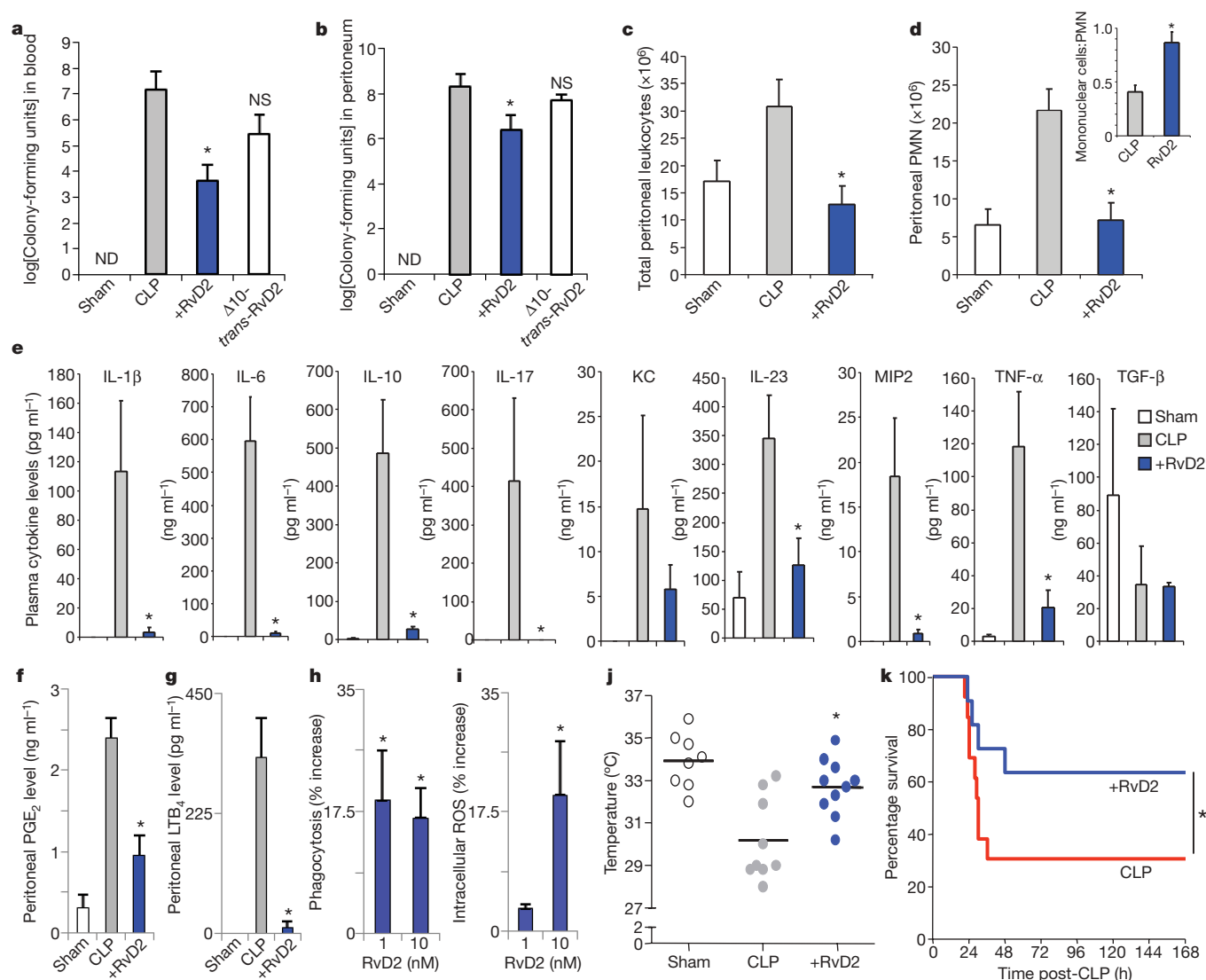


Figure 4 | RvD2 reduces bacterial levels, systemic inflammation and enhances survival in microbial sepsis. **a, b**, Aerobic bacteria levels in blood (**a**) and peritoneal exudates (**b**) from sham or CLP-operated mice \pm RvD2 methyl ester (RvD2-Me) (100 ng; intravenously) at 12 h ($n = 5-6$ per group). ND, not detected; NS, not significant. **c, d**, Peritoneal leukocyte differentials, ratio of mononuclear to polymorphonuclear cells inset. ($n = 3-5$). **e**, Plasma

cytokine levels ($n = 3-6$). **f, g**, Peritoneal PGE₂ and LTB₄ levels. **h, i**, Phagocytosis of *E. coli* and intracellular ROS in human PMN. **j**, Temperatures of mice 12 h post-CLP. **k**, Kaplan-Meier survival analysis of vehicle- ($n = 13$) and RvD2-treated ($n = 12$) CLP mice. Results are mean \pm s.e.m. **a-g**, * $P < 0.05$ two-tailed unpaired Student's *t*-test. **h, j**, * $P < 0.05$ by one-way ANOVA. **k**, * $P < 0.05$ one-tailed log-rank test.

is plausible that RvD2 prevents persistent amplification signals downstream of pattern-recognition receptors, dampening the responses of classically activated macrophages^{23,24}.

RvD2 also drastically decreased IL-17, as well as IL-10, which is of interest because of its detrimental impact on survival in sepsis²⁵. Thus, RvD2 differs in action from lipoxin A₄, which stimulates IL-10 (ref. 26). High levels of both pro- and anti-inflammatory cytokines, including IL-10, are predictive of early mortality in sepsis, and diminishing IL-10 levels proved beneficial in sepsis^{27,28}. Pro-inflammatory mediators, including prostaglandin E₂ (PGE₂) and LTB₄, were also decreased in peritoneum by RvD2 (Fig. 4f, g and Supplementary Fig. 10b, c). Interestingly, in addition to macrophage-directed actions, RvD2 directly enhanced PMN *Escherichia coli* phagocytosis that was accompanied by an increase in intracellular ROS (Fig. 4h, i). RvD2 does not possess direct antibacterial activity compared to ampicillin (Supplementary Fig. 11). RvD2-treated CLP mice also showed protection at 12 h post-CLP from hypothermia (Fig. 4j). Accordingly, RvD2 dramatically increased survival rates among CLP-operated mice (Fig. 4k) and activity levels 12 h post-CLP were resumed (Supplementary movie 3).

The present results establish the complete stereochemistry of endogenous RvD2 and its potent stereoselective actions facilitating resolution. Local and systemic bacterial burden in microbial sepsis were controlled and significantly dampened with RvD2. This potent D-series resolvins protected CLP-mice from excessive leukocyte infiltration and overzealous cytokine production, and also enhanced clearance of microbes, thus preventing sepsis-induced lethality. Sepsis remains a clinical challenge, with high mortality rates and increasing prevalence^{5,29}. Given the uncontrolled inflammatory pathogenesis of sepsis, anti-inflammatory therapies are used for sepsis management in humans, but have ultimately failed owing primarily to sustained immunosuppression⁵. Overall, these results indicate that RvD2 is a potent endogenous mediator which actively promotes the resolution of inflammation, suggesting new therapeutic approaches that do not compromise host defence.

METHODS SUMMARY

Intra-vital microscopy. Experiments were approved by and performed under guidelines of the Ethical Committee for Use of Animals, Barts and The London School of Medicine and Home Office regulations (Guidance on the Operation of

Animals, Scientific Procedures Act, 1986). Intra-vital microscopy was used to observe actions of RvD2 on leukocyte responses stimulated by PAF (100 nM C16 form C₂₆H₅₄NO₇P; Sigma; PAF administered 1 h before RvD2 was given at time 0) within the cremasteric microcirculation of C57 BL/6 mice. Mice were anaesthetized with xylazine (7.5 mg kg⁻¹) and ketamine (150 mg kg⁻¹), and cremaster prepared³⁰. In some experiments, fluorescein isothiocyanate (FITC)-albumin (1 mg) was administered intravenously to assess vascular leakage.

Caecal ligation and puncture. CLP was performed in male FVB mice¹⁷, in accordance with the Harvard Medical Area Standing Committee On Animals Protocol #02570. The caecum was ligated below the ileocaecal valve for mid-grade sepsis¹⁷. A through and through puncture was performed with a 20-gauge needle, followed by one additional puncture in the distal tip of the caecum. Mice received saline (500 µl subcutaneously) followed by intravenous administration of vehicle (0.1% ethanol), or RvD2 methyl ester (100 ng) at the time of puncture. In some experiments, RvD2-Me (1 µg) was administered intraperitoneally 1 h post-CLP. At 12 h, rectal temperature was measured, blood collected by cardiac puncture and peritoneal exudates obtained. Blood and peritoneal bacteria levels were determined by growth on tryptic soy agar plates. Plasma and peritoneal cytokine levels were determined by Searchlight array. Peritoneal cells were differentiated using Wright-Giemsa staining.

Statistics. Data are mean ± s.e.m. Multiple group comparisons were made using one-way or two-way analysis of variance (ANOVA) followed by Dunnett's or Bonferroni post tests where appropriate and direct comparisons made using a two-tailed unpaired Student's *t*-test. Kaplan–Meier survival curves were analysed using a one-tailed log-rank test. In all cases, a *P* value <0.05 was considered significant.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 August; accepted 21 September 2009.

- Gilroy, D. W., Lawrence, T., Perretti, M. & Rossi, A. G. Inflammatory resolution: new opportunities for drug discovery. *Nature Rev. Drug Discov.* **3**, 401–416 (2004).
- Serhan, C. N., Chiang, N. & Van Dyke, T. E. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nature Rev. Immunol.* **8**, 349–361 (2008).
- Serhan, C. N. *et al.* Resolvins: a family of bioactive products of omega-3 fatty acid transformation circuits initiated by aspirin treatment that counter proinflammation signals. *J. Exp. Med.* **196**, 1025–1037 (2002).
- Weber, C., Zernecke, A. & Libby, P. The multifaceted contributions of leukocyte subsets to atherosclerosis: lessons from mouse models. *Nature Rev. Immunol.* **8**, 802–815 (2008).
- Hotchkiss, R. S. & Karl, I. E. The pathophysiology and treatment of sepsis. *N. Engl. J. Med.* **348**, 138–150 (2003).
- Shimizu, T. Lipid mediators in health and disease: enzymes and receptors as therapeutic targets for the regulation of immunity and inflammation. *Annu. Rev. Pharmacol. Toxicol.* **49**, 123–150 (2009).
- Hong, S., Gronert, K., Devchand, P. R., Moussignac, R. L. & Serhan, C. N. Novel docosatrienes and 17S-resolvins generated from docosahexaenoic acid in murine brain, human blood, and glial cells. Autacoids in anti-inflammation. *J. Biol. Chem.* **278**, 14677–14687 (2003).
- Shimizu, T., Radmark, O. & Samuelsson, B. Enzyme with dual lipoxygenase activities catalyzes leukotriene A₄ synthesis from arachidonic acid. *Proc. Natl Acad. Sci. USA* **81**, 689–693 (1984).
- Rodríguez, A. R. & Spur, B. W. First total synthesis of 7(S),16(R),17(S)-Resolvin D₂, a potent anti-inflammatory lipid mediator. *Tetrahedr. Lett.* **45**, 8717–8720 (2004).
- Prescott, S. M., Zimmerman, G. A., Stafforini, D. M. & McIntyre, T. M. Platelet-activating factor and related lipid mediators. *Annu. Rev. Biochem.* **69**, 419–445 (2000).
- Cooper, D., Norling, L. V. & Perretti, M. Novel insights into the inhibitory effects of Galectin-1 on neutrophil recruitment under flow. *J. Leukoc. Biol.* **83**, 1459–1466 (2008).
- Rittirsch, D. *et al.* Functional roles for C5a receptors in sepsis. *Nature Med.* **14**, 551–557 (2008).
- Kubes, P., Suzuki, M. & Granger, D. N. Nitric oxide: an endogenous modulator of leukocyte adhesion. *Proc. Natl Acad. Sci. USA* **88**, 4651–4655 (1991).
- Moncada, S. & Higgs, E. A. The discovery of nitric oxide and its role in vascular biology. *Br. J. Pharmacol.* **147** (Suppl. 1), S193–S201 (2006).
- Bucci, M. *et al.* Endothelial nitric oxide synthase activation is critical for vascular leakage during acute inflammation *in vivo*. *Proc. Natl Acad. Sci. USA* **102**, 904–908 (2005).
- Paul-Clark, M. J., Van Cao, T., Moradi-Bidhendi, N., Cooper, D. & Gilroy, D. W. 15-epi-lipoxin A₄-mediated induction of nitric oxide explains how aspirin inhibits acute inflammation. *J. Exp. Med.* **200**, 69–78 (2004).
- Rittirsch, D., Huber-Lang, M. S., Flierl, M. A. & Ward, P. A. Immunodesign of experimental sepsis by cecal ligation and puncture. *Nature Protocols* **4**, 31–36 (2009).
- Buras, J. A., Holzmann, B. & Sitkovsky, M. Animal models of sepsis: setting the stage. *Nature Rev. Drug Discov.* **4**, 854–865 (2005).
- Singer, P. *et al.* Anti-inflammatory properties of omega-3 fatty acids in critical illness: novel mechanisms and an integrative perspective. *Intensive Care Med.* **34**, 1580–1592 (2008).
- Farolan, L. R., Goto, M., Myers, T. F., Anderson, C. L. & Zeller, W. P. Perinatal nutrition enriched with omega-3 polyunsaturated fatty acids attenuates endotoxin shock in newborn rats. *Shock* **6**, 263–266 (1996).
- Pluess, T. T. *et al.* Intravenous fish oil blunts the physiological response to endotoxin in healthy subjects. *Intensive Care Med.* **33**, 789–797 (2007).
- Haworth, R. *et al.* The macrophage scavenger receptor type A is expressed by activated macrophages and protects the host against lethal endotoxin shock. *J. Exp. Med.* **186**, 1431–1439 (1997).
- Litvak, V. *et al.* Function of C/EBPΔ in a regulatory circuit that discriminates between transient and persistent TLR4-induced signals. *Nature Immunol.* **10**, 437–443 (2009).
- Mosser, D. M. & Edwards, J. P. Exploring the full spectrum of macrophage activation. *Nature Rev. Immunol.* **8**, 958–969 (2008).
- Flierl, M. A. *et al.* Adverse functions of IL-17A in experimental sepsis. *FASEB J.* **22**, 2198–2205 (2008).
- Souza, D. G. *et al.* The required role of endogenously produced lipoxin A₄ and annexin-1 for the production of IL-10 and inflammatory hyporesponsiveness in mice. *J. Immunol.* **179**, 8533–8543 (2007).
- Osuchowski, M. F., Welch, K., Siddiqui, J. & Remick, D. G. Circulating cytokine/inhibitor profiles reshape the understanding of the SIRS/CARS continuum in sepsis and predict mortality. *J. Immunol.* **177**, 1967–1974 (2006).
- Huang, X. *et al.* PD-1 expression by macrophages plays a pathologic role in altering microbial clearance and the innate inflammatory response to sepsis. *Proc. Natl Acad. Sci. USA* **106**, 6303–6308 (2009).
- Dombrovskiy, V. Y., Martin, A. A., Sunderram, J. & Paz, H. L. Rapid increase in hospitalization and mortality rates for severe sepsis in the United States: a trend analysis from 1993 to 2003. *Crit. Care Med.* **35**, 1244–1250 (2007).
- Chatterjee, B. E. *et al.* Annexin 1-deficient neutrophils exhibit enhanced transmigration *in vivo* and increased responsiveness *in vitro*. *J. Leukoc. Biol.* **78**, 639–646 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge support from National Institutes of Health grants GM-38765 and P50-DE016191 (C.N.S.), Wellcome Trust Programme grant 086867/Z/08/Z (R.J.F. and M.P.) and Project grant 085903/Z/08 (R.J.F.) and Arthritis Research Campaign UK fellowships 18445 and 18103 (to L.V.N. and D.C., respectively). M.S. received a National Research Service Award from the NHLBI (HL087526). We thank J. W. Winkler and J. Uddin for work related to RvD2 synthesis, P. Pillai, K. Martinod, G. Fredman and J. Dalli for technical assistance, and M. H. Small for assistance with the manuscript. We also thank B. Schmidt for histopathology, Children's Hospital Boston.

Author Contributions M.S. and L.V.N. designed and carried out experiments, analysed data and wrote the manuscript; L.S., R.Y. and D.C. carried out experiments and analysed data; N.A.P. synthesized RvD2; R.J.F. and M.P. designed experiments, analysed data and contributed to the manuscript; C.N.S. planned the project, designed experiments, analysed data and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.N.S. (cnsrhan@zeus.bwh.harvard.edu).

METHODS

RvD2 total organic synthesis. Total organic synthesis of RvD2 will be reported elsewhere. Briefly, stereochemically pure RvD2 methyl ester was prepared using chiral starting materials having the same stereochemistry of the three OH groups and using stereochemically controlled processes (Supplementary Fig. 1a). Thus, the synthesis started with the nucleophilic opening of enantiomerically pure protected glycidol **1** with a protected derivative of pentynoic acid **2**, followed by functional group manipulations to form intermediate **3**. Synthesis of the intermediate that contains the 16,17-diol moiety was prepared starting with a protected form of 2-deoxyribose **4**, which was converted to intermediate **5** using a Wittig reaction that selectively formed a C = C bond with a Z-geometry. Conversion of **5** to alkyne intermediate **6**, followed by palladium-mediated Sonogashira coupling of **6** with iodide **3** gave compound **7**, the acetylenic precursor of RvD2. Removal of the protective groups and selective hydrogenation to convert the alkyne moiety to a C = C bond with a Z-geometry gave the RvD2 methyl ester (RvD2-Me). This compound was purified by high-performance liquid chromatography (HPLC) and fully characterized with NMR spectroscopy and mass spectrometry (Supplementary Fig. 1 and Fig. 1c and f). This is a different synthetic strategy for RvD2 than reported in ref. 9. The $\Delta 10$ -*trans*-RvD2 was prepared by isomerization at room temperature (21 °C) and exposure to light overnight, and was isolated using reverse phase (RP)-HPLC before bioassay.

RvD2 biosynthesis. Isolated human PMNs were isolated from healthy volunteers (BWH protocol # 88-02462) suspended in DPBS+/+ (25×10^6 cells ml⁻¹) and incubated with 5 μ g 17S-hydroperoxy-4Z,7Z,10Z,13Z,15E,19Z-docosahexaenoic acid (17-HpDHA) and/or 17-HDHA and stimulated with zymosan (100 μ g) or A23187 (5 μ M) for 30 min at 37 °C (ref. 3). Incubations were stopped with cold methanol and samples were solid-phase extracted³¹.

Mediator lipidomics. LC/UV/MS/MS-based mediator lipidomic analysis was performed with an HPLC (Shimadzu LC20AD) connected inline with a UV diode array detector (Agilent G1315B), coupled to a hybrid quadrupole time-of-flight mass spectrometer (QStar XL; Applied Biosystems) equipped with a Phenomenex Luna C18(2) column (2 mm \times 150 mm \times 3 μ m). Acquisition was carried out in negative ionization mode. The mobile phase consisted of methanol/water/acetic acid (60:40:0.01, v/v/v) and was ramped to 85:15:0.01 over 30 min and to 100:0:0.01 over the next 5 min at a flow rate of 200 μ l min⁻¹. The flow rate was increased to 400 μ l min⁻¹ over the next 10 min. The operating parameters were: gas1 35.00, gas2 55.00, curtain gas 40, electrospray voltage -4,200 V, evaporation temperature 400 °C, decluster potential -35 V, and collision energies were optimized individually. GC/MS analysis was carried out with an HP 6890 gas chromatography system equipped with an HP-5MS capillary column (0.25 mm inner diameter \times 30 m, Agilent) and an HP5973 mass selective detector (Hewlett-Packard)³¹.

Targeted LC-MS/MS-based lipidomics of CLP exudates. Exudate lavages (2 ml) were collected and immediately added to 2 volumes of cold methanol for extraction and work-up as in ref. 31. LC-MS/MS identification was acquired with an Agilent 1100 series HPLC paired with an ABI Sciex Instruments 3200 Q TRAP linear ion trap quadrupole mass spectrometer. The column (Agilent Eclipse Plus C18, 4.6 mm \times 50 mm \times 1.8 μ m) was eluted at a flow rate of 0.4 ml min⁻¹ with methanol/water/acetic acid (60/40/0.01;v/v/v) ramped to 80/20/0.01 (v/v/v) after 5 min, 95/5/0.01 (v/v/v) after 8 min, and 100/0/0.01 after 14 min to wash the column. Instrument control and data acquisition were performed using Analyst 1.4.2 software. Ion pair transitions from previously reported multiple reaction monitoring methods were used for profiling and quantitation of PGE₂ (351.2/189.2) and LTB₄ (335.2/195.2). Criteria for identification were: liquid chromatography retention time and a minimum of six fragment diagnostic ions on the MS/MS spectrum matching those of synthetic standards. Deuterated 5S-HETE (Cayman Chemicals, 2 ng) was added before extraction as internal standard for recovery calculations.

Murine peritonitis and macrophage phagocytosis. Peritonitis was assessed using male FVB mice (Charles River) or eNOS^{-/-} mice and their wild-type littermates (C57BL/6J; Jackson Labs). Vehicle (1.0% EtOH) or RvD2 (0.001–100 ng per mouse) were administered intravenously followed by intraperitoneal

administration of zymosan A (1 mg; Sigma). Peritoneal lavages were collected after 4 h and leukocyte infiltration was assessed by light microscopy, followed by differential analysis using Wright–Giemsa staining. For macrophage phagocytosis, resident peritoneal macrophages were harvested from naive mice and incubated with RvD2 (0.01–100 nM) for 15 min at 37 °C before the addition of opsonized FITC-labelled zymosan. After 30 min, extracellular fluorescence was quenched by trypan blue and fluorescence was determined using a Victor3 plate reader (PerkinElmer).

Direct actions of RvD2 on human endothelial cells and PMN. Primary HUVECs (Lonza) were incubated with or without pertussis toxin (100 ng ml⁻¹, 12 h) with vehicle alone (0.1% EtOH) or RvD2 (0.01–100.0 nM) for 30 min at 37 °C and supernatants were analysed for NO_x (nitrate or nitrite) by the Greiss reaction (Biomol) and 6-keto PGF_{1 α} by ELISA (Neogen). Human PMN were isolated from healthy donors by Ficoll gradient and incubated with PAF (1 nM) with or without RvD2 (0.01–10 nM) for 15 min at 37 °C and the surface expression of CD62L and CD18 was determined by flow cytometry using phycoerythrin-conjugated anti-CD18 (6.7) and FITC-conjugated anti-CD62L (Dreg 56; BD Biosciences) or appropriate isotype control (IgG1 κ). To assess leukocyte–endothelial interactions, the flow chamber assay was used¹¹, human PMN were stimulated before flow with PAF (1 nM) or C5a (0.1 nM) for 15 min.

Extracellular superoxide assay. PMN were incubated with 5 μ M lucigenin in a thermostatted (37 °C) luminometer (Wallac VICTOR2 1420 Multilabel Counter, Perkin Elmer). After 15 min, PMN were treated with vehicle, C5a (10 nM) or RvD2 (1–10 nM) and monitored for 20 min. Alternatively, PMN were incubated with RvD2 (1–10 nM) for 15 min before vehicle or C5a (10 nM). **Assessment of *in vivo* vascular responses by fluorescence imaging.** Male FVB mice (6–8 weeks) were anaesthetized with pentobarbital (50 mg kg⁻¹, intraperitoneally) and injected intravenously with FITC-albumin (2.5 mg) alone or plus L-NAME (24 mg kg⁻¹). Acetone or RvD2-methyl ester (20 μ l at indicated doses) was applied to the inner ear. Fluorescence was assessed using the Night Owl LB 981 NC 320 Molecular Light Imager (Berthold Technologies) and images were acquired for 1,500 ms (every 5 min) using a HQ 475 excitation filter (Chroma) and Winlight Software. The mean change in fluorescence intensity was normalized to acetone control. In some experiments, RvD2 (10 pg; intravenously) was administered in conjunction with FITC-albumin.

In separate experiments, topical application of LTB₄ (1.0 μ g; Cayman), RvD2-Me (0.1 or 0.01 ng) or acetone were applied to the inner ear. After 24 h, 6 mm punch biopsies were obtained, flash frozen in liquid N₂, stored at -80 °C and myeloperoxidase activity was assessed as in ref. 32.

***E. coli* phagocytosis and generation of intracellular ROS.** Adherent human PMN were pre-incubated with RvD2 (0.1–100 nM) or vehicle for 15 min before incubation with *E. coli* (JM109; 50:1 ratio) at 37 °C for 60 min. PMN were rinsed, fixed and permeabilized (BD Cytotfix/Cytoperm; BD Biosciences), and *E. coli* levels were assessed using FITC-conjugated anti-*E. coli* antibody (GTX40856; GeneTex) by flow cytometry. To assess intracellular ROS generation, PMN were incubated with 5 μ M carboxy-H₂DCFDA (C400; Invitrogen) for 30 min before incubation with RvD2 and *E. coli*, and probe oxidation was determined using a Victor3 plate reader (Perkin Elmer).

Antibacterial susceptibility test. Mid-logarithmic phase *E. coli* (5×10^7 colony-forming units) was inoculated on LB agar plates, discs containing RvD2 (0.1–100 ng) or ampicillin (10 μ g) were placed on top, and the zone of clearing was assessed after overnight incubation.

Immunohistochemistry. Inguinal lymph nodes and spleens were excised 12 h after CLP and fixed in 3% formalin. Tissues were paraffin-embedded, sectioned and slides were stained with haematoxylin and eosin and Gram by the Department of Pathology at the Children's Hospital Boston.

- Serhan, C. N., Lu, Y., Hong, S. & Yang, R. Mediator lipidomics: search algorithms for eicosanoids, resolvins, and protectins. *Methods Enzymol.* **432**, 275–317 (2007).
- Takano, T., Clish, C. B., Gronert, K., Petasis, N. & Serhan, C. N. Neutrophil-mediated changes in vascular permeability are inhibited by topical application of aspirin-triggered 15-epi-lipoxin A4 and novel lipoxin B4 stable analogues. *J. Clin. Invest.* **101**, 819–826 (1998).

LETTERS

Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells

Siqin Bao^{1*}, Fuchou Tang^{1*}, Xihe Li², Katsuhiko Hayashi^{1†}, Astrid Gillich¹, Kaiqin Lao³ & M. Azim Surani¹

The pluripotent state, which is first established in the primitive ectoderm cells of blastocysts, is lost progressively and irreversibly during subsequent development¹. For example, development of post-implantation epiblast cells from primitive ectoderm involves significant transcriptional and epigenetic changes, including DNA methylation and X chromosome inactivation², which create a robust epigenetic barrier and prevent their reversion to a primitive-ectoderm-like state. Epiblast cells are refractory to leukaemia inhibitory factor (LIF)–STAT3 signalling, but they respond to activin/basic fibroblast growth factor to form self-renewing epiblast stem cells (EpiSCs), which exhibit essential properties of epiblast cells^{3,4} and that differ from embryonic stem (ES) cells derived from primitive ectoderm⁵. Here we show reprogramming of advanced epiblast cells from embryonic day 5.5–7.5 mouse embryos with uniform expression of N-cadherin and inactive X chromosome to ES-cell-like cells (rESCs) in response to LIF–STAT3 signalling. Cultured epiblast cells overcome the epigenetic barrier progressively as they proceed with the erasure of key properties of epiblast cells, resulting in DNA demethylation, X reactivation and expression of E-cadherin. The accompanying changes in the transcriptome result in a loss of phenotypic and epigenetic memory of epiblast cells. Using this approach, we report reversion of established EpiSCs to rESCs. Moreover, unlike epiblast and EpiSCs, rESCs contribute to somatic tissues and germ cells in chimaeras. Further studies may reveal how signalling-induced epigenetic reprogramming may promote reacquisition of pluripotency.

Previous studies showed that epiblast cells, unlike primitive ectoderm cells, are refractory to LIF–STAT3 signalling; instead they respond to activin/basic fibroblast growth factor (bFGF) to generate EpiSCs^{3,4}, which are more like epiblast with an inactive X chromosome, and differ from ES cells. Here we re-examined advanced post-implantation epiblast cells to see if they could revert to ES-cell-like cells in response to LIF–STAT3 signalling.

First, epiblast tissue carrying an Oct4–ΔPE–green fluorescent protein (GFP) reporter⁶ was isolated from mouse embryos on embryonic days (E) E5.5–E7.5. This reporter, with only the distal enhancer for *Oct4* (also known as *Pou5f1*), shows preferential expression in the primitive ectoderm, primordial germ cells (PGCs) and ES cells, but not in the epiblast or EpiSCs⁶. Notably, the distal enhancer constitutes an ‘enhanceosome’ representing the densest binding locus for the key pluripotency-specific transcripts in ES cells⁷, which makes it likely that its activation will only occur if all pluripotency factors are expressed optimally; some of these must be lacking or suboptimal in the epiblast and EpiSCs.

Next, the epiblast tissue was dissected to remove the most proximal region (the site of existing PGCs and PGC precursors²) and the outer visceral endoderm (Fig. 1a). Notably, unlike previous studies

where the epiblast tissue was left intact^{3,4}, we trypsin-digested individual epiblast into a single-cell suspension to break up existing cell–cell interactions and promote establishment of a new signalling-induced transcriptional network *in vitro*. The resulting cells were cultured in LIF–fetal calf serum (FCS) on mouse embryonic fibroblast (MEF) feeder cells, a standard condition used for the derivation of ES cells from primitive ectoderm, and of induced pluripotent stem (iPS) cells from somatic cells^{5,8,9}. After 4–7 days, most cultures revealed large colonies (Fig. 1a and Supplementary Table 1) with many alkaline-phosphatase-positive cells (Fig. 1a), but no detectable Oct4–ΔPE–GFP expression, indicating that the distal enhancer was yet inactive (Fig. 1a). We could propagate the cultured epiblast (cEpi) colonies in LIF–FCS after collagenase treatment without detectable changes for at least 20 passages.

After culture of cEpi cells for 14–35 days, we started to detect clusters of GFP-positive cells within cEpi colonies (Fig. 1b, c), indicating activation of the distal enhanceosome of the Oct4–ΔPE–GFP reporter⁶. Subsequent culture of GFP-positive cells was carried out after disruption of cEpi colonies by treatment with trypsin, which is detrimental for cEpi cells but promotes propagation of ES-cell-like cells. With further passaging, we established ES-cell-like cells with uniform GFP expression (Fig. 1b). We call these cells reprogrammed epiblast ES-cell-like cells (rESCs).

The frequency of rESC derivation was relatively high at around 22–36%, which notably did not diminish with developmental age from E5.5 to E7.5 (Supplementary Table 1). Furthermore, the epiblast cells at the start were uniformly negative for Oct4–ΔPE–GFP expression and positive for N-cadherin and inactive X-chromosome (see below). If reversion had occurred from rare epiblast cells, we might have seen a reduction in the frequency of rESC derivation. Notably, no ES-cell-like cells have previously been reported from embryos as late as E7.5; rather, they have only been derived from primitive ectoderm present in the inner cell mass in blastocysts^{5,7,9}. Furthermore, pluripotent EPL¹⁰ and FAB-SC¹¹ were derived from pre-implantation or implanting blastocysts and not from post-implantation embryos (Supplementary Table 2); their epigenetic state was not reported.

To gain an insight into the reversion process, we examined changes in gene expression (Fig. 2a, b). Transcriptome analysis revealed that cEpi cells, like EpiSCs^{3,4}, were closely related to the epiblast. Thus, cEpi cells showed strong expression of *Eomes*, *Fgf5*, *Sox17*, *Gata6*, *Lefty1* and *Cer1*. However, there was little expression of *stella* (also called *Dppa3*), *Pecam1*, *Rex1* (also called *Zfp42*) and *Fbxo15*, but their expression increased in rESCs with a concomitant loss of *Fgf5*, *Eomes* and *Sox17* (Fig. 2a). Expression of key pluripotency genes *Oct4*, *Sox2* and *Nanog* also increased in rESCs (Fig. 2a and Supplementary Fig. 1, 2b). Expression of *Eomes* and *Fgf5* was slightly higher in early passage rESCs (passage 4 (P4)) compared to the levels in P24 cells

¹Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QN, UK. ²College of Life Science, Inner Mongolia University/Mengniu RB CO. Ltd., West No. 1 Daxue Road, Huhhot, Inner Mongolia 010021, China. ³Molecular Cell Biology, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404, USA. [†]Present address: Department of Anatomy and Cell Biology, Graduate School of Medicine, Kyoto University, Yoshida-Konoe-Cho, Sako-Ku, Kyoto 606-8501, Japan.

*These authors contributed equally to this work.

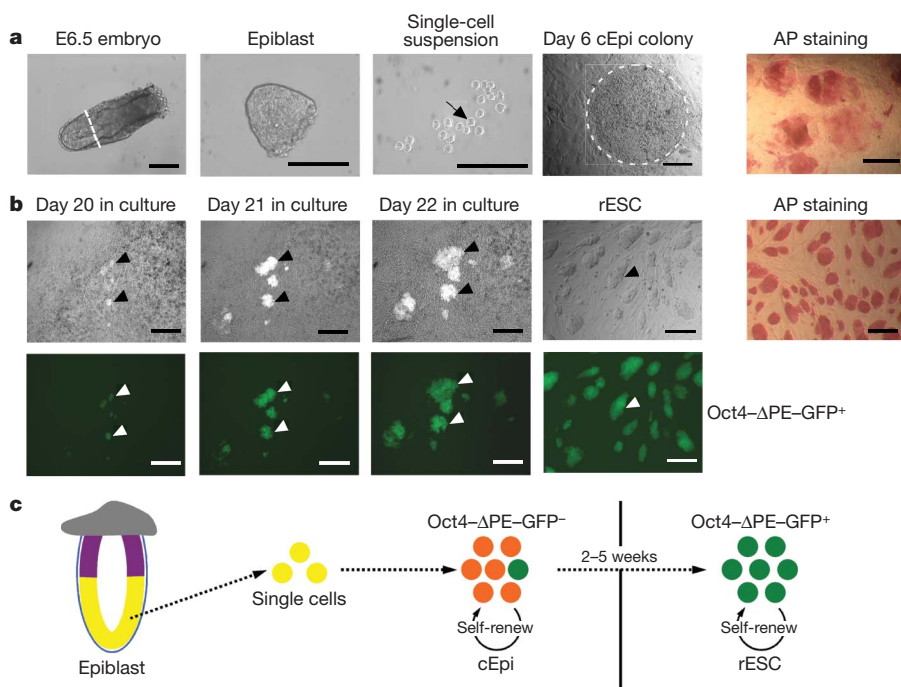


Figure 1 | Reprogramming epiblast cells from mouse E6.5 embryos to generate rESCs.

a, Derivation of cEpi cells from E6.5 epiblast. Epiblast tissue had the proximal region (white line, far left panel) and visceral endoderm removed; a single-cell suspension (black arrow) was cultured, which formed cEpi colonies. Note alkaline phosphatase (AP)-positive cells in cEpi colonies in the far right panel. **b**, Derivation of rESCs from cEpi colonies. Note the appearance of clusters of Oct4-ΔPE-GFP-positive cells in cEpi colonies (black arrowheads), and corresponding white arrowheads for GFP in the panel below. Scale bars, 100 μm (**a** and **b**). **c**, Schematic representation of reprogramming of epiblast through cEpi cells, and finally rESCs in response to LIF-STAT3 signalling.

(Fig. 2a), suggesting a loss of residual ‘memory’ of their epiblast origin (see below). Thus, comprehensive microarray analysis confirmed that rESCs are similar to ES cells and differ from cEpi cells and EpiSCs (Fig. 2b and Supplementary Fig. 3). These changes in the transcriptome are consistent with the distal enhancer-driven activation of the Oct4-ΔPE-GFP reporter in rESCs.

We next asked if LIF-STAT3 signalling is critical for reprogramming, and found that STAT3 is phosphorylated in cEpi cells, suggesting that they respond to LIF signalling (Supplementary Fig. 4a). Notably, addition of the Janus protein tyrosine kinase inhibitor (JAK inhibitor 1; Calbiochem) that prevents phosphorylation of tyrosine 705 of STAT3 initially allowed some cEpi colonies to develop, but they gradually differentiated and failed to form rESCs (Supplementary Table 3). Thus, LIF-STAT3 seems to be crucial for the propagation and reprogramming of cEpi cells to rESCs. Furthermore, culture of rESCs with the JAK inhibitor caused a marked reversal towards the cEpi-like transcriptome (Supplementary Fig. 4b). A number of STAT3 targets and their expression have been identified in ES cells⁷, including *Fbxo15*, *Rex1* and *Stat3* itself, as well as the epigenetic modifiers *Lin28*, *Ezh2* and *Mbd3*, suggesting that STAT3 has the potential to influence the transcriptional and epigenetic state of cEpi cells.

Next, we examined epigenetic changes in the epiblast during reversion to rESCs. Notably, reactivation of the late-replicating inactive X chromosome^{12,13} during reversion to rESCs would indicate a major epigenetic change¹⁴. Consistently, we found that all the E6.5 epiblast cells (96 of 96 cells; Supplementary Fig. 5a) had the characteristic accumulation of histone H3 lysine 27 trimethylation (H3K27me3) associated with the inactive X chromosome (ref. 15). After 12 days, nearly all cEpi cells (99 of 100 cells; Fig. 3a) still had the H3K27me3 ‘spot’, which declined to 62% after 25 days (37 of 60 cells), and to only 9% after 35 days of culture (7 of 80 cells). This suggests continuing epigenetic reprogramming towards rESCs, which uniformly lacked the H3K27me3 ‘spot’. These observations on H3K27me3 suggest initiation of X reactivation, a hallmark of epigenetic reprogramming, although further studies are needed to confirm completion of the process. Similar changes are also seen after somatic nuclear transplantation into oocytes, in mouse iPS cells, and in ES-cell-somatic cell hybrids^{16–19}.

We also examined DNA methylation of the promoter regions of *stella* and *Rex1*; both of these genes (and others like *Pecam1*) are

repressed in the epiblast but active in primitive ectoderm and ES cells²⁰. Although *stella* and *Rex1* were unmethylated in the epiblast, they became transiently methylated in cEpi cells before undergoing demethylation (Fig. 3b), which is consistent with the activation of the *Stella*-GFP reporter in rESCs (Supplementary Fig. 2a). Once established, the rESC epigenotype was stable, heritable and distinct from EpiSCs, which retain key properties of epiblast cells, including inactive X, and in which *stella* and *Rex1* are methylated and repressed (Figs 2a and 3b). These findings are relevant to human ES cells, which resemble mouse EpiSCs but not mouse ES cells or rESCs.

To observe the dynamic nature of reprogramming to rESCs, we also examined changes in the expression of E-cadherin and N-cadherin (Fig. 4a). Whereas expression of both E-cadherin and N-cadherin was detected in E6.5 epiblast uniformly, trypsin-digestion of these cells before culture led to the loss of these adhesion molecules. During subsequent culture, we detected heterogeneous N-cadherin expression in cEpi cells, but in continuing culture there was a complete loss of N-cadherin, which was replaced by uniform expression of E-cadherin in rESCs, consistent with the evidence that LIF-STAT3 promotes upregulation of E-cadherin¹¹.

Next, we asked if rESCs might have originated from early PGCs because they can undergo dedifferentiation into pluripotent embryonic germ cells that are similar to ES cells. To reduce this likelihood, the epiblast tissue from E6.5–E7.5 mouse embryos was dissected away from the most proximal region, the site of PGC precursors and PGC, respectively. In particular, E7.5 epiblast also shows a loss of competence to form additional PGCs²¹. Second, PGCs require bFGF and stem cell factor for proliferation and dedifferentiation into pluripotent embryonic germ cells, and cannot survive in culture conditions used for cEpi cells. Furthermore, rESCs, unlike embryonic germ cells, retain methylation of imprinted genes^{22,23} (Supplementary Fig. 5c). Thus, cumulative evidence makes it unlikely that rESCs could originate from PGCs.

In view of our observations, we asked if established EpiSCs cultured in activin/bFGF could also undergo reversion to rESCs in response to LIF-STAT3. We chose two EpiSC lines: one with the Oct4-ΔPE-GFP reporter (passage 20) and the other with an X-GFP reporter; the latter was FACS sorted to establish lines with repressed reporter on the inactive X chromosome (passage 23) (Supplementary Fig. 5b). When these EpiSCs with stably repressed GFP reporters were cultured for 10–20 days (Supplementary Fig. 6a, b), we detected GFP-positive

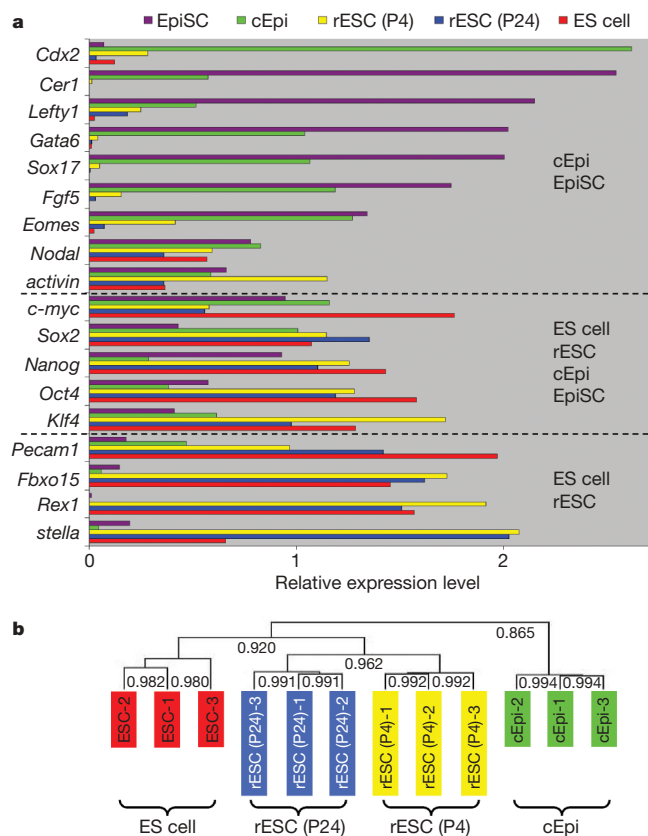


Figure 2 | Changes in gene expression profile. **a**, Reverse transcription real-time PCR of marker genes in EpiSCs, cEpi cells and rESCs at early (P4) and late (P24) passages, and ES cells. Note progressive loss of markers of epiblast detected in cEpi cells and EpiSCs (at the top) and enhancement of expression of genes in rESCs that resemble those expressed in ES cells. **b**, Whole-genome cluster analysis of transcriptomes of cEpi cells, rESCs at early (P4) and late (P24) passages, and ES cells. The numbers correspond to Pearson correlation coefficients between different cDNA samples. Note that rESCs resemble ES cells but not cEpi cells, which are more like the original epiblast cells as described above.

cells, from which we established rESCs as before. Re-expression of the X-GFP reporter indicates X reactivation. Furthermore, we found that the stably repressed *stella* and *Rex1* became de-repressed after DNA demethylation during reversion to rESCs (Figs 2a and 3b). Thus, unlike previous studies^{24,25}, we found that reversion of EpiSCs is possible in response to LIF-STAT3 without exogenous transcription factors. We had previously shown that ES cells in LIF-FCS exist in a metastable state and fluctuate between ES cell and epiblast-like states but without proceeding completely to the EpiSC-like state. Because the presence of feeder cells apparently helps to promote a shift towards an ES-cell-like state²⁰, it is possible that they help to promote the reprogramming process.

Finally, we tested the developmental potential of rESCs in chimaeric embryos. Using rESCs derived from epiblast with the ROSA-LacZ reporter, we found an extensive contribution to developing embryos and in adults with germline transmission (Supplementary Fig. 7 and Supplementary Tables 4–6). Although early passage rESCs contributed to the extra-embryonic ectoderm (a trophectoderm derivative) in E6.5 embryos (Supplementary Fig. 7b), this was not seen with late-passage rESCs (Supplementary Table 6 and Supplementary Fig. 7c). Thus, a transient memory of the epiblast origin in rESCs is lost with progressive changes in the transcriptome and the epigenetic state. Most notably, rESCs from EpiSCs could also participate in chimaeras and contribute to the germ line in E13.5 embryos (Supplementary Fig. 7i and Supplementary Table 5), which is not possible with

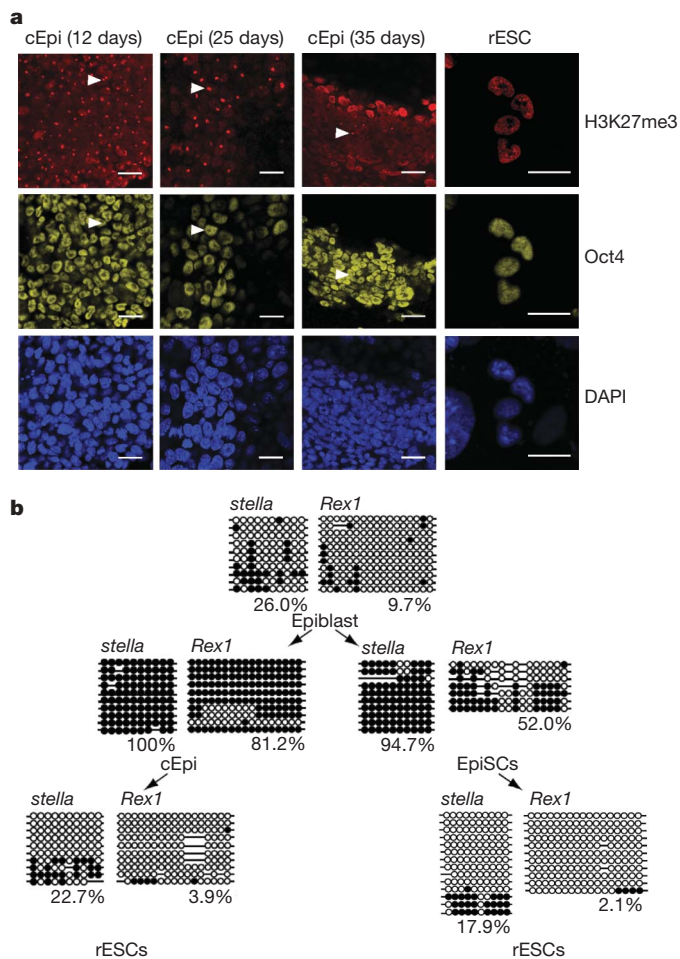


Figure 3 | Epigenetic changes during reprogramming of epiblast cells. **a**, Female cEpi cells exhibit uniform accumulation of H3K27me3 associated with the inactive X (white arrowhead), which is gradually lost from individual cells during culture and finally lost in rESCs. Scale bar, 20 μ m. DAPI, 4,6-diamidino-2-phenylindole. **b**, Changes in DNA methylation of *stella* and *Rex1* during reprogramming of epiblast. Although *stella* and *Rex1* are repressed in epiblast cells, these loci are initially unmethylated; they undergo DNA methylation transiently in cEpi cells and stably in EpiSCs. Reprogramming to form rESCs results in loss of DNA methylation.

EpiSCs^{3,4}, indicating that rESCs derived from EpiSCs undergo a stable reversion to an ES-cell-like state.

Post-implantation epiblast cells and established EpiSCs can ‘overcome’ a robust epigenetic barrier and undergo reversion to rESCs in response to LIF-STAT3 (Fig. 4b). This observation is significant for human ES cells, which resemble mouse EpiSCs^{3,4,26}. Further studies may provide critical insights into signal-induced epigenetic reprogramming, including DNA demethylation and X-reactivation, which also represent a major barrier during reprogramming of somatic cells to iPS cells^{9,20,27,28}. Notably, specification of PGCs from epiblast is also accompanied by similar epigenetic reprogramming events by a different mechanism involving the expression of key germ-cell determinants, including *Blimp1/Prdm1* and *Prdm14* (refs 2, 21); these genes repress the somatic program and initiate epigenetic reprogramming during specification of PGCs from epiblast cells²⁹. As a result, PGCs are epigenetically similar to primitive ectoderm cells (Fig. 4b, refs 29–31). Epigenetic changes in PGCs, and during the experimental reversion of epiblast and EpiSCs to rESCs, may provide novel insights into diverse mechanisms underlying epigenetic reprogramming of epiblast cells, and contribute generally to the understanding of the role of epigenetic mechanisms in other contexts.

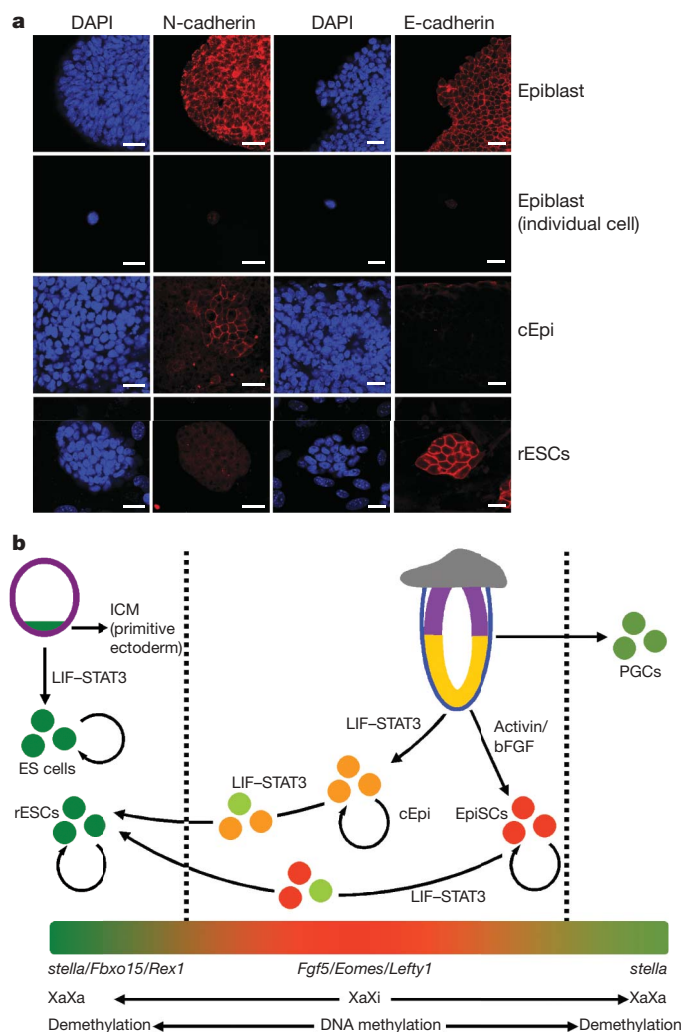


Figure 4 | Dynamic changes of cell surface markers and model of reprogramming. **a**, Dynamic changes in cell surface adhesion molecules. Both E-cadherin and N-cadherin are detected uniformly in E6.5 epiblast, which is undetectable in single-cell suspension. During culture, N-cadherin expression is heterogeneous in cEpi cells and eventually disappears completely and is replaced by E-cadherin in rESCs. Scale bars, 20 μ m. **b**, Schematic representation of reprogramming of 'epiblast' cEpi cells and EpiSCs to rESCs. Note the epigenetic and transcriptional changes during reprogramming of cEpi cells and EpiSCs. Specification of primordial germ cells (PGC) also results in epigenetic reprogramming, including expression of *stella* and X reactivation. ICM, inner cell mass; Xa, active X chromosome; Xi, inactive X chromosome.

METHODS SUMMARY

Post-implantation epiblast from mouse E5.5–E7.5 embryos was isolated by cutting out the extra-embryonic and proximal epiblast cells with glass needles. The epiblast was then treated with EGTA and trypsin, and dissociated into single cells by pipetting with a hand-pulled glass capillary. The single-cell suspension from individual epiblast tissue was cultured in standard ES cell medium with LIF and FCS on feeder cells. These cells formed colonies called cultured epiblast cells (cEpi), and were regularly passaged on feeder cells at 3–6-day intervals. After culture of cEpi cells for 14–35 days, about 10–50 GFP-positive cells appeared in individual cEpi colonies. When these GFP-positive cell clusters grew to 100–200 μ m diameter, they were treated with trypsin, and the resulting cells were cultured to produce GFP-positive colonies. We call these cells rESCs.

Received 8 December 2008; accepted 28 September 2009.
Published online 8 October 2009.

- Gardner, R. L. & Rossant, J. Investigation of the fate of 4–5 day post-coitum mouse inner cell mass cells by blastocyst injection. *J. Embryol. Exp. Morphol.* **52**, 141–152 (1979).

- Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128**, 747–762 (2007).
- Tesar, P. J. et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).
- Brons, I. G. et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191–195 (2007).
- Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).
- Yeom, Y. I. et al. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonic cells. *Development* **122**, 881–894 (1996).
- Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
- Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
- Niwa, H. How is pluripotency determined and maintained? *Development* **134**, 635–646 (2007).
- Rathjen, J., Washington, J. M., Bettess, M. D. & Rathjen, P. D. Identification of a biological activity that supports maintenance and proliferation of pluripotent cells from the primitive ectoderm of the mouse. *Biol. Reprod.* **69**, 1863–1871 (2003).
- Chou, Y. F. et al. The growth factor environment defines distinct pluripotent ground states in novel blastocyst-derived stem cells. *Cell* **135**, 449–461 (2008).
- Takagi, N., Sugawara, O. & Sasaki, M. Regional and temporal changes in the pattern of X-chromosome replication during the early post-implantation development of the female mouse. *Chromosoma* **85**, 275–286 (1982).
- Silva, J. et al. Establishment of histone h3 methylation on the inactive X chromosome requires transient recruitment of Eed-Enx1 polycomb group complexes. *Dev. Cell* **4**, 481–495 (2003).
- Chuva de Sousa Lopes, S. M. et al. X chromosome activity in mouse XX primordial germ cells. *PLoS Genet.* **4**, e30 (2008).
- Plath, K. et al. Role of histone H3 lysine 27 methylation in X inactivation. *Science* **300**, 131–135 (2003).
- Mikkelsen, T. S. et al. Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55 (2008).
- Maherali, N. et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1**, 55–70 (2007).
- Tada, M., Takahama, Y., Abe, K., Nakatsuiji, N. & Tada, T. Nuclear reprogramming of somatic cells by *in vitro* hybridization with ES cells. *Curr. Biol.* **11**, 1553–1558 (2001).
- Bao, S. et al. Initiation of epigenetic reprogramming of the X chromosome in somatic nuclei transplanted to a mouse oocyte. *EMBO Rep.* **6**, 748–754 (2005).
- Hayashi, K., Lopes, S. M., Tang, F. & Surani, M. A. Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* **3**, 391–401 (2008).
- Ohinata, Y. et al. A signaling principle for the specification of the germ cell lineage in mice. *Cell* **137**, 571–584 (2009).
- Shovlin, T. C., Durcova-Hills, G., Surani, A. & McLaren, A. Heterogeneity in imprinted methylation patterns of pluripotent embryonic germ cells derived from pre-migratory mouse germ cells. *Dev. Biol.* **313**, 674–681 (2008).
- Meissner, A. et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
- Guo, G. et al. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063–1069 (2009).
- Hanna, J. et al. Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell* **4**, 513–524 (2009).
- Thomson, J. A. et al. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147 (1998).
- Silva, J. et al. Promotion of reprogramming to ground state pluripotency by signal inhibition. *PLoS Biol.* **6**, e253 (2008).
- Wernig, M. et al. A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nature Biotechnol.* **26**, 916–924 (2008).
- Hayashi, K. & Surani, M. A. Resetting the epigenome beyond pluripotency in the germline. *Cell Stem Cell* **4**, 493–498 (2009).
- Surani, M. A., Durcova-Hills, G., Hajkova, P., Hayashi, K. & Tee, W. W. Germ line, stem cells, and epigenetic reprogramming. *Cold Spring Harb. Symp. Quant. Biol.* **73**, 9–15 (2008).
- Hayashi, K. & Surani, M. A. Self-renewing epiblast stem cells exhibit continual delineation of germ cells with epigenetic reprogramming *in vitro*. *Development* **136**, 3549–3556 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Lee for assistance. This work was supported by grants from the Wellcome Trust to M.A.S.

Author Contributions S.B., F.T., X.L. and M.A.S. designed the research project; S.B. and F.T. performed most of the experiments, with contributions from X.L., K.H. and A.G.; microarray analysis was performed by K.L.; S.B., F.T., K.H., A.G. and M.A.S. carried out critical assessment of the data; M.A.S. wrote the paper with input from all the authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.A.S. (a.surani@gurdon.cam.ac.uk).

LETTERS

A regulatory circuit for *piwi* by the large Maf gene *traffic jam* in *Drosophila*

Kuniaki Saito¹, Sachi Inagaki¹, Toutai Mituyama², Yoshinori Kawamura^{1,3}, Yukiteru Ono⁴, Eri Sakota⁵, Hazuki Kotani¹, Kiyoshi Asai^{2,6}, Haruhiko Siomi¹ & Mikiko C. Siomi^{1,7}

PIWI-interacting RNAs (piRNAs) silence retrotransposons in *Drosophila* germ lines by associating with the PIWI proteins Argonaute 3 (AGO3), Aubergine (Aub) and Piwi^{1,2}. piRNAs in *Drosophila* are produced from intergenic repetitive genes and piRNA clusters by two systems: the primary processing pathway and the amplification loop^{1–7}. The amplification loop occurs in a Dicer-independent, PIWI-Slicer-dependent manner^{3,4,8}. However, primary piRNA processing remains elusive. Here we analysed piRNA processing in a *Drosophila* ovarian somatic cell line where Piwi, but not Aub or AGO3, is expressed; thus, only the primary piRNAs exist. In addition to *flamenca*, a Piwi-specific piRNA cluster³, *traffic jam* (*tj*)⁹, a large Maf gene, was determined as a new piRNA cluster. piRNAs arising from *tj* correspond to the untranslated regions of *tj* messenger RNA and are sense-oriented. piRNA loading on to Piwi may occur in the cytoplasm. *zucchini*¹⁰, a gene encoding a putative cytoplasmic nuclease, is required for *tj*-derived piRNA production. In *tj* and *piwi* mutant ovaries, somatic cells fail to intermingle with germ cells and Fasciclin III is overexpressed. Loss of *tj* abolishes Piwi expression in gonadal somatic cells. Thus, in gonadal somatic cells, *tj* gives rise simultaneously to two different molecules: the TJ protein, which activates Piwi expression, and piRNAs, which define the Piwi targets for silencing.

Genetic studies have shown that *piwi* and *aub* are essential in germline stem-cell self-renewal and pole-cell formation, respectively^{11,12}. Mutations introduced into *piwi* and *aub* cause de-repression of retrotransposons and a loss of piRNA accumulation in ovaries⁸. A recent study has revealed that strong loss-of-function mutations in *AGO3* also increase expression of selfish genetic elements in germ lines⁵. Thus, the PIWI proteins, with their associated piRNAs, function in retrotransposon silencing. piRNA production in *Drosophila* ovaries occurs in a Dicer-independent manner⁸. A model for piRNA biogenesis—the piRNA amplification loop^{3,4}—was proposed as a result of deep-sequencing and bioinformatic analyses of *Drosophila* piRNAs. In this model, Aub/Piwi and AGO3 reciprocally guide the 5' end formation of piRNAs.

Classification of piRNAs according to their origins indicated that piRNAs derived from a particular piRNA cluster locus—*flamenca* (*flam*)—on the X chromosome are exclusively loaded on to Piwi^{1,3}, indicating that those piRNAs are produced by a pathway independent of the amplification loop^{1,2}. This pathway is called the primary processing pathway^{1,2}. Recently, two independent groups deduced the existence of the primary piRNA processing pathway from extensive bioinformatic analyses of piRNAs in a broad range of piRNA mutants; these studies reconfirmed that primary piRNAs derived from *flam* are most likely loaded directly on to Piwi and not further amplified^{5,6}.

However, a molecular mechanism of the primary processing pathway remains elusive.

We established a stable cell line of ovarian somatic cells (OSCs) from the parental cell line fGS/OSS, comprising germline stem cells and sheets of somatic cells (OSS)¹³. fGS/OSS culture was shown to be Vasa-positive¹³, whereas our OSCs are Vasa-negative (Supplementary Fig. 1a). OSCs express Fasciclin III (FasIII; Supplementary Fig. 1b) and undergo rounds of passage in culture for several months. All these data support the idea that OSCs contain only mitotically active early follicle (somatic) cells.

Piwi is expressed in somatic gonadal cells^{3,14,15}, whereas Aub and AGO3 are not expressed in this cell type^{3,4,16}. Western blot analysis revealed that Piwi, but not Aub and AGO3, was detectable in OSCs (Supplementary Fig. 1a). As in ovaries, Piwi in OSCs was localized in the nucleus (Supplementary Fig. 1c). The absence of AGO3 expression in OSCs was further confirmed by polymerase chain reaction with reverse transcription (RT-PCR; Supplementary Fig. 1d, e).

Piwi in OSCs was detected in a form bound to small RNAs of 24–30 nucleotides (Fig. 1a). The size distribution of these small RNAs was very similar to that of gonadal Piwi-associating piRNAs (Piwi piRNAs)^{3,4,14}. In addition, they showed resistance to periodate oxidation and β -elimination treatments, which are hallmarks of 2'-O-methyl modification at the 3' end, as in ovarian piRNAs^{8,17,18} (Supplementary Fig. 1f). Thus, the small RNAs associated with Piwi in OSCs could be categorized as piRNAs and, importantly, these piRNAs are produced in an Aub/AGO3-independent manner.

We examined whether the nuclear localization of Piwi is required for piRNA production and for Piwi loading in OSCs. A mutant of Piwi (Piwi- Δ N) in which its putative nuclear localization signals were deleted, resulting in cytoplasmic localization of the mutant (Fig. 1b), was loaded with piRNAs in a similar manner to wild-type Piwi (Fig. 1c). We also observed that a Slicer-deficient Piwi mutant—Piwi-DDAA, where two aspartic acids (D614 and D685) in the PIWI domain, which are required for Slicer activity, are altered to alanines—was loaded with piRNAs, similarly to wild-type Piwi (Fig. 1c). Depletion of endogenous Piwi from OSCs did not affect piRNA loading on to a double mutant of Piwi (Piwi- Δ N-DDAA) (Supplementary Fig. 1g). Piwi does not seem to homodimerize *in vivo* (data not shown). These results support a model in which the primary piRNA processing and the piRNA loading on Piwi may occur in the cytoplasm in a Piwi-Slicer-independent manner.

Piwi-associating piRNAs in OSCs (OSC piRNAs) are mainly derived from the antisense strand of retrotransposons (Fig. 2a and Supplementary Fig. 2a), similar to the derivation of ovarian Piwi piRNAs^{3,4,14,19}. The size distribution of OSC piRNAs is shown in

¹Keio University School of Medicine, Tokyo 160-8582, Japan. ²Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan. ³Institute of Health Biosciences University of Tokushima, Tokushima 770-8503, Japan. ⁴Department of Life Sciences, Information and Mathematical Science Laboratory, Inc., Tokyo 112-0012, Japan. ⁵Japan Biological Informatics Consortium (JBIC), Tokyo 135-8073, Japan. ⁶Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8583, Japan. ⁷JST, CREST, Saitama 332-0012, Japan.

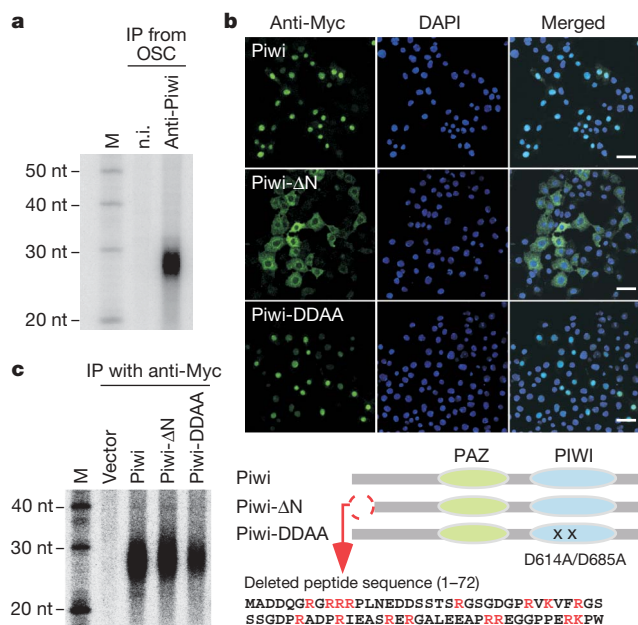


Figure 1 | Piwi in OSCs is associated with endogenous small RNAs. **a**, Piwi-associating small RNAs in OSCs were visualized by 32 P-labelling. n.i., non-immune IgG was used as a negative control. **b**, Subcellular localization of Myc-tagged wild-type Piwi, Piwi-ΔN and Piwi-DDAA in OSCs. Scale bars, 20 μ m. DAPI, 4', 6-diamidino-2-phenylindole. **c**, Myc-tagged wild-type Piwi, Piwi-ΔN and Piwi-DDAA expressed in OSCs are bound with piRNAs. M, molecular mass marker.

Supplementary Fig. 2b. Examination of their nucleotide bias indicated that OSC piRNAs mostly have a bias for U as the first nucleotide in the sequence (1st-U), but no other prominent bias was observed throughout their entire sequence (Fig. 2b). Exclusion of piRNAs with 1st-U from the piRNA pool did not uncover any obvious bias, including 10th-A (Supplementary Fig. 2c). piRNA pairings through the ten nucleotides from the 5' ends were negligible (Supplementary Fig. 2d). Thus, Piwi does not self-amplify piRNAs. All these observations correlate well with the data obtained from a recent deep-sequencing study that was performed using a *Drosophila* ovarian somatic sheet (OSS)⁷.

The origins of OSC piRNAs were examined (Fig. 2c and Supplementary Fig. 3). Unique mapping (see Methods for definition) of OSC piRNAs on the *Drosophila* genome revealed that *flam* is the main source (1,365 perfectly matched and 186 one-base mismatched piRNAs; Supplementary Figs 3 and 4), as it is for ovarian Piwi piRNAs³ and OSS piRNAs⁷. In addition to *flam*, we found another locus on chromosome 2L that also produces piRNAs uniquely mapped to the particular region (322 perfectly matched and 29 one-base mismatched piRNAs; Fig. 2c; chromosome 2L, unique). This locus corresponds to the 3' untranslated region (UTR) of a protein-coding, single-exon gene, *traffic jam*⁹ (*tj*, Fig. 2d).

TJ is a soma-specific large Maf factor necessary for controlling gonad morphogenesis in *Drosophila*⁹. TJ is the only *Drosophila* orthologue of the transcriptional factors c-Maf and MafB/Kleisler in vertebrates⁹. In *tj* mutant gonads, somatic cells fail to intermingle and properly develop germ cells. This eventually causes an early blockage in germ-cell differentiation and no follicle cells are detected in adult ovaries of *tj* mutants⁹. All of the OSC piRNAs derived from *tj* were sense-oriented (Fig. 2d), indicating that the *tj* transcript may serve not only as the template for TJ synthesis but also as the precursor of the piRNAs. The *tj* transcript in OSCs was not dedicated to piRNA production, because the TJ protein was strongly expressed in OSCs (Supplementary Fig. 5a). The existence of piRNAs derived from the 3' UTR of *tj* was further confirmed by northern blot analysis (Supplementary Fig. 5b). *tj*-derived piRNAs also appeared in ovarian total

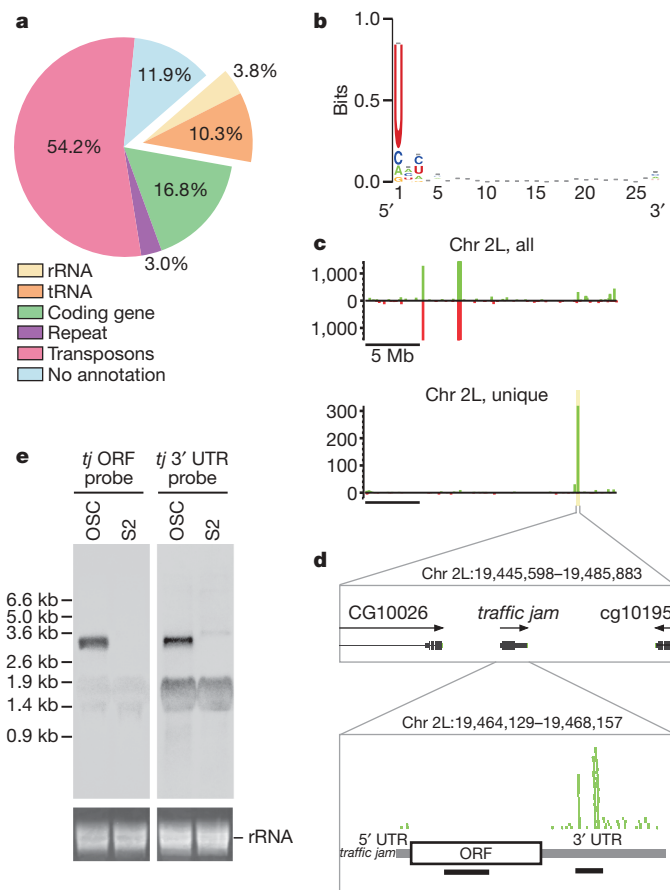


Figure 2 | Piwi-associated piRNAs in OSCs. **a**, The contents of Piwi-associated small RNAs in OSCs perfectly match the *Drosophila* genome sequence. **b**, Examination of nucleotide bias indicates that Piwi piRNAs have mostly uracil at the 5' ends (1st-U). **c**, Mapping data on chromosome 2L. Unique mapping data (bottom; shown as 'unique') indicate that *tj* is the source of OSC piRNAs. y axis shows the read number. **d**, The locus containing *tj* (shown in **c**) is enlarged. Green bars (bottom) indicate piRNAs corresponding to *tj*. Two thick black lines show probes (ORF and 3' UTR) in **e**. **e**, Northern blot analysis. Both ORF and 3' UTR probes visualized a single band of the same length corresponding to the *tj* transcript in OSCs.

small RNA libraries produced by ref. 20 (Supplementary Fig. 6) and ref. 6 (see later).

We assessed whether the transcriptional unit of *tj* is first divided into two parts, each with an individual function—one for TJ synthesis and the other for piRNA production—or if one full-length *tj* transcriptional unit contains both functions. Northern blot analysis, using two probes corresponding to either the open reading frame (ORF) or the 3' UTR of *tj*, visualized a single discrete band of the same length (Fig. 2e), indicating that the latter scenario might be the case.

piRNAs derived from 3' UTRs (and from slightly extended regions) of protein-coding genes other than *tj* were also found (Supplementary Table 1). For example, the 3' UTRs of *brat*²¹ and *Klp10A*²² also produce piRNAs (Supplementary Fig. 7). Interestingly, they are all derived from the sense strand; thus the parental genes are apparently not the targets for gene silencing by Piwi. The parental genes are also not repetitive. By contrast, piRNAs derived from retrotransposons or *flam* in OSCs are mainly antisense-oriented and are thought to arise from much longer, repetitive precursors¹.

Mutations in *zucchini* (*zuc*), a gene encoding a putative nuclease, cause female sterility¹⁰ and a reduction of *roo*- and *flam*-derived piRNAs in ovaries^{6,10}. However, whether or not *zuc* is expressed in gonadal somatic cells and, if it is expressed, where *Zuc* accumulates was unknown. We observed that OSCs express *zuc* (data not shown)

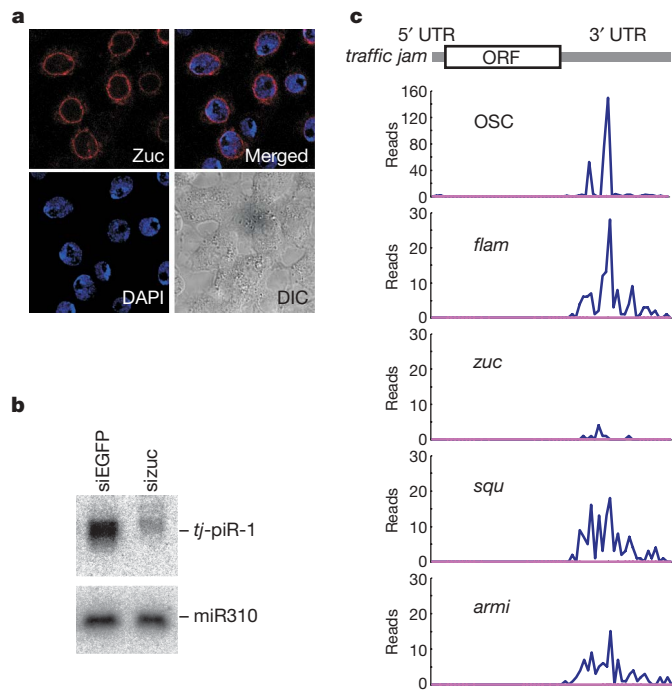


Figure 3 | Involvement of *zuc* in the *tj*-piRNA production pathway. **a**, Myc-tagged Zuc protein was overexpressed in OSCs and its subcellular localization was determined using an anti-Myc antibody (red). Zuc accumulates in the perinuclear region in the cytoplasm of OSCs. DIC, differential interference contrast. **b**, Accumulation of *tj*-derived piRNAs, but not of miRNAs, requires *zuc* in OSCs. **c**, The read numbers of *tj*-derived piRNAs in various piRNA mutants. These data, obtained from the data set of ref. 6, strongly support the idea that *zuc* is required for *tj*-piRNA production.

and that Zuc is predominantly localized in the cytoplasm of OSCs, particularly in the perinuclear region (Fig. 3a). We then sought to determine whether *zuc* is necessary for *tj* piRNA production in OSCs. Depletion of *zuc* by RNA interference (RNAi) significantly reduced the expression level of *tj* piRNAs, but not of microRNAs (Fig. 3b), suggesting that *zuc* is involved in the *tj* piRNA production pathway. This was further supported by analysis of the data set of ref. 6, which indicated that the read number of *tj* piRNAs in *zuc* mutants was much lower compared with those in other mutants (Fig. 3c).

It was previously reported that *tj* mutant somatic cells show a failure to intermingle with germ cells in third instar larval ovaries⁹. Notably, we noticed that the *piwi* mutants²³ *piwi*² and *piwi*³ showed a similar phenotype: somatic cells in the ovaries of third instar larvae of *piwi* mutants adhere to each other and exclude Vasa-positive germ cells (Fig. 4a and Supplementary Fig. 8a). *aub* mutants did not phenocopy the *piwi* mutants (Supplementary Fig. 8b). *piwi* mutants express the *tj* transcript (data not shown) and the TJ protein (Fig. 4a and Supplementary Fig. 8a) at approximately wild-type levels. Thus it is unlikely that *tj* piRNAs target the parental *tj* gene.

We next examined the expression of Piwi in larval ovaries of *tj* mutants. Vasa and TJ are known to be expressed in germline stem cells (and in their developing cells) and somatic cells, respectively, whereas Piwi is known to be expressed in both cell types^{3,11,14}. In wild-type larval ovaries, Vasa- and TJ-positive cells are mutually exclusive, but Piwi is expressed in both cell populations (Fig. 4b and Supplementary Fig. 9a). By contrast, in *tj* mutants, Piwi expression was restricted to Vasa-positive cells (Fig. 4b and Supplementary Fig. 9a). These results indicate that TJ is the activator of *piwi* expression in gonadal somatic cells.

Expression of Piwi in *tj* mutant testes was also examined. As in ovaries, *tj* mutations in testes caused a lack of Piwi expression in gonadal somatic cells, including hub, cyst progenitor cells and early cyst cells. In contrast, Piwi signals were clearly seen in these cell types of wild-type testes (Supplementary Fig. 9b). Interestingly, without functional TJ, germline stem cells and their developing cells in testes, which normally show a faint signal for Piwi, highly express Piwi (Supplementary Fig. 9b). It seems that TJ in testes negatively controls

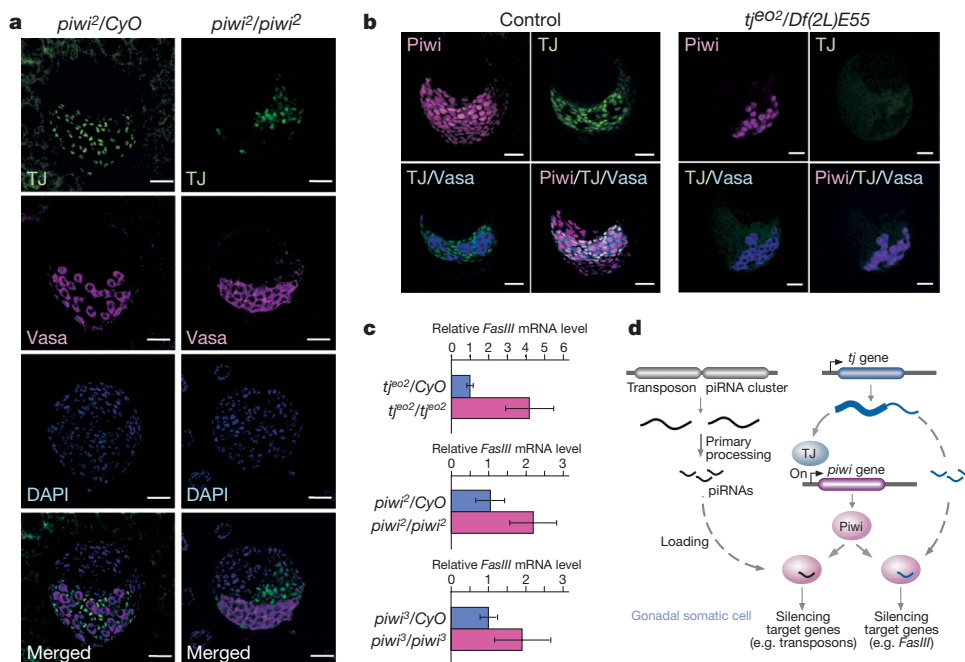


Figure 4 | Phenotypes of *tj* and *piwi* mutant ovaries and testes. **a**, In control (*piwi*²/*CyO*) ovaries, TJ-positive cells and primordial germ cells (PGCs) form a mixed cell population. In *piwi*²/*piwi*² larval ovaries, TJ-positive cells do not intermingle with PGCs, but instead form coherent clusters. TJ and Vasa are shown in green and magenta, respectively. Scale bars, 20 μ m. **b**, In *tj* mutant (*tj*^{eo2}/*Df*(2L)*E55*) larval ovaries, Piwi (magenta)

is expressed only in Vasa-positive (blue) cells. TJ is shown in green. Scale bars, 20 μ m. **c**, Quantitative RT-PCR shows that the expression level of *FasIII* is upregulated by loss of *piwi* or *tj* expression. **d**, Two functions of *tj* in the regulation of the function of *piwi* in gonadal somatic cells. The first function of *tj* is to activate *piwi* expression. The second is to supply piRNAs for Piwi. Genes targeted by the complex might include *FasIII*.

piwi expression and indirectly controls expression in germline stem cells and their developing cells.

This study has uncovered two functions of *tj* in the regulation of *piwi*'s functions. In gonadal somatic cells, TJ supposedly controls transcription of various genes. Our study indicates that *piwi* is highly likely to be one of the genes under strong TJ control because loss of *tj* in gonadal somatic cells abolished Piwi expression. Further support for the hypothesis that TJ controls *piwi* expression was provided by DNA sequences near the putative transcriptional start site of the *piwi* gene, which show a weak but significant similarity to the Maf binding consensus sequence²⁴, and which were bound with TJ in OSCs (Supplementary Fig. 10a, b). Thus, the first function of *tj* is to activate the expression of Piwi in gonadal somatic cells. The second function is to supply piRNAs for Piwi. Without the supplement of *tj* piRNAs, Piwi would lose the activity to target genes that should be silenced by Piwi and the *tj*-piRNA complex. A likely target of such silencing is *FasIII* because *FasIII*, a cell adhesion molecule concentrated at the hub cell junction, is ectopically overexpressed in other somatic cells in *tj* larval testes⁹. Indeed, the *FasIII* expression level was higher in *piwi* mutant testes than in control testes (Fig. 4c). Some of the *tj* piRNAs identified in this study showed strong complementarity to the *FasIII* primary transcript (Supplementary Fig. 10c). Although all the *tj*-derived piRNAs are sense-oriented and thus unlikely to target the *tj* mRNA, given the nuclear localization of Piwi, it is conceivable that the Piwi-piRNA complex could associate with the *tj* gene.

These findings suggest a novel regulatory circuit where *tj* mRNA simultaneously produces two types of functional molecules (Fig. 4d): TJ protein, which activates expression of Piwi, and piRNAs, which are loaded on to Piwi to silence specific target genes, such as *FasIII* and other, as yet undiscovered, genes.

METHODS SUMMARY

The OSC line was developed from fGS/OSS¹³. Piwi was immunopurified from OSCs using an anti-Piwi antibody¹⁴. Cloning of small RNAs associated with Piwi in OSCs was carried out as described²⁵. Genome mapping and annotation was performed as described in the Methods. Western blotting⁴, RT-PCR, peroxidation/ β -elimination^{4,17}, northern blotting¹⁴ and immunostaining²⁶ were performed as described.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 August; accepted 16 September 2009.

Published online 7 October 2009.

- Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761–764 (2007).
- Siomi, H. & Siomi, M. C. On the road to reading the RNA-interference code. *Nature* **457**, 396–404 (2009).
- Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
- Gunawardane, L. S. et al. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* **315**, 1587–1590 (2007).
- Li, C. et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**, 509–521 (2009).
- Malone, C. et al. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137**, 522–535 (2009).
- Lau, N. et al. Abundant primary piRNAs, endo-siRNAs and microRNAs in a *Drosophila* ovary cell line. *Genome Res.* doi:10.1101/gr.094896.109 (14 July 2009).
- Vagin, V. V. et al. A distinct small RNA pathway selfish genetic elements in the germline. *Science* **313**, 320–324 (2006).

- Li, M. A., Alls, J. D., Avancini, R. M., Koo, K. & Godt, D. The large Maf factor Traffic Jam controls gonad morphogenesis in *Drosophila*. *Nature Cell Biol.* **5**, 994–1000 (2003).
- Pane, A., Wehr, K. & Schupbach, T. *zucchini* and *squash* encode two putative nucleases required for rasiRNA production in the *Drosophila* germline. *Dev. Cell* **12**, 851–862 (2007).
- Cox, D. N. et al. A novel class of evolutionarily conserved genes defined by *piwi* are essential for stem cell self-renewal. *Genes Dev.* **12**, 3715–3727 (1998).
- Harris, A. N. & Macdonald, P. M. *aubergine* encodes a *Drosophila* polar granule component required for pole cell formation and related to eIF2C. *Development* **128**, 2823–2832 (2001).
- Niki, Y., Yamaguchi, T. & Mahowald, A. P. Establishment of stable cell lines of *Drosophila* germ-line stem cells. *Proc. Natl Acad. Sci. USA* **103**, 16325–16330 (2006).
- Saito, K. et al. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev.* **20**, 2214–2222 (2006).
- Cox, D. N., Chao, A. & Lin, H. *piwi* encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Genes Dev.* **127**, 503–514 (2000).
- Nishida, M. K. et al. Gene silencing mechanisms mediated by Aubergine piRNA complexes in *Drosophila* male gonad. *RNA* **13**, 1911–1922 (2007).
- Saito, K. et al. Pimet, the *Drosophila* homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends. *Genes Dev.* **21**, 1603–1608 (2007).
- Horwich, M. D. et al. The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr. Biol.* **17**, 1265–1272 (2007).
- Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in *Drosophila melanogaster*. *Nature* **450**, 304–308 (2007).
- Czech, B. et al. An endogenous small interfering RNA pathway in *Drosophila*. *Nature* **453**, 798–802 (2008).
- Arama, E., Dickman, D., Kimchie, Z., Shearn, A. & Lev, Z. Mutations in the beta-propeller domain of the *Drosophila* brain tumor (*brat*) protein induce neoplasm in the larval brain. *Oncogene* **19**, 3706–3716 (2000).
- Rogers, G. C. et al. Two mitotic kinesins cooperate to drive sister chromatid separation during anaphase. *Nature* **427**, 364–370 (2004).
- Lin, H. & Spradling, A. C. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development* **124**, 2463–2476 (1997).
- Kataoka, K. Multiple mechanisms and functions of Maf transcriptional factors in the regulation of tissue-specific genes. *J. Biochem.* **141**, 775–781 (2007).
- Kawamura, Y. et al. *Drosophila* endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature* **453**, 793–797 (2008).
- Kitadate, Y., Shigenobu, S., Arita, K. & Kobayashi, S. Boss/Sev signaling from germline to soma restricts germline-stem-cell-niche formation in the anterior region of *Drosophila* male gonads. *Dev. Cell* **13**, 151–159 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Niki, D. Godt, H. Lin, E. Matunis, S. Kobayashi, Y. Kitadate, Y. Kageyama and H. Sano for providing reagents. We also thank Bloomington and Kyoto *Drosophila* Stock Center for the supply of *Drosophila* strains. We thank K. Yamada, E. Hattori, K. M. Nishida and T. N. Okada for technical assistance; S. Takahashi and K. Kataoka for discussions and suggestions; and other members of the Siomi laboratory for discussions and comments on the manuscript. We also thank K. Greer and D. McGowan for encouragement. This work was supported by MEXT grants to H.S. and NEDO (New Energy and Industrial Technology Development Organization) grants to M.C.S., T.M. and K.A. M.C.S. is supported by CREST from JST. M.C.S. is Associate Professor of Global COE for Human Metabolomics Systems Biology by MEXT.

Author Contributions K.S., S.I., Y.K. and H.K. conducted biochemical experiments. T.M., Y.O., E.S. and K.A. performed bioinformatics. K.S., T.M., H.S. and M.C.S. designed experiments, interpreted data and prepared the manuscript.

Author Information Small RNA sequences have been deposited at the GEO database under accession number GSE15137. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.S. (awa403@sc.itc.keio.ac.jp) or M.C.S. (siomim@sc.itc.keio.ac.jp).

METHODS

OSC culture. Culture medium was prepared essentially as previously described¹³. Cross and Sang's M3 (BF) medium was prepared from Shields and Sang M3 Insect Medium supplemented with 0.6 mg ml⁻¹ glutathione, 10% FBS, 10 mU ml⁻¹ insulin and 10% fly extract. We collected adult Oregon R flies and measured their total weight. Flies were treated with 70% ethanol for 5 min, and then washed several times with PBS. Flies were transferred to a mortar and 10 ml of M3 medium supplemented with 10% FBS (M3/FBS) was added. Flies were homogenized using a pestle on ice, after which M3/FBS was added to the homogenate, giving a final concentration of 0.2 g (fly weight) per ml. This homogenate was centrifuged at 6,000g for 10 min. The supernatant was divided into 500 µl aliquots and then heat treated at 60 °C for 5 min. This preparation was spun at 17,000g for 10 min. The supernatant was centrifuged at least twice more and the clear supernatant then collected. The supernatant (fly extract) was stored at -30 °C before use. Fly extract was sterilized by filtration and added to the culture medium to a final concentration of 10%. For passaging, OSCs were washed in PBS and then exposed to a solution of trypsin-EDTA for 1 min at 37 °C. Culture medium was then added to neutralize the trypsin. The cells were washed by aspiration in the medium and aliquots were pipetted into a 6-well plate (Falcon) or a 90-mm cell culturing dish (FPI). These were then incubated at 26 °C (the detailed protocol is available on request).

Drosophila strains. *yellow white* (*y w*) and Oregon R were used as wild-type strains. The *piwi* alleles used were *piwi*² and *piwi*³ (Bloomington stock number 12225) (gift from H. Lin). *Df(2L)BSC145* (Bloomington stock number 9505) is a deficiency line that covers the *piwi* locus. The *tj* alleles used were *tj*⁹⁰² and *tj*⁴⁷³⁵ (gift from D. Golt, S. Kobayashi and H. Sano). *Df(2L)E55* (DGRC stock number 106822) is a deficiency line that covers the *tj* locus. The *aub* alleles used were *aub*^{HN2} *cn bw/CyO* and *aub*^{QC42} *cn bw/CyO*. *aub*^{HN2} *cn bw/CyO* and *aub*^{QC42} *cn bw/CyO* were crossed to yield *aub* heterozygous flies, *aub*^{QC42}/*aub*^{HN2}. To recognize *piwi*^{-/-}, *tj*^{-/-} and *aub*^{QC42}/*aub*^{HN2} larvae, *w*^{*}; *L*² *Pin*¹/*CyO*, *P{GAL4-Kr.C}DC3 P{UAS-GFP.S65T}DC7* (Bloomington stock number 5194) were used as balancer chromosomes. All stocks were maintained at 25 °C.

Immunoprecipitation and cloning of small RNAs associated with Piwi in OSCs. Piwi was immunopurified from OSCs using a specific antibody¹⁴. 1 × 10⁸ OSCs were homogenized in a hypotonic buffer (30 mM HEPES, pH 7.3, 2 mM magnesium acetate, 5 mM dithiothreitol (DTT) and 1 mg ml⁻¹ Pefablock SC) to prepare OSC lysate. Piwi and AGO1 were immunopurified using anti-AGO1 (ref. 27) and anti-Piwi (P3G11) antibodies¹⁴ immobilized on GammaBind beads (GE Healthcare). Just before immunoprecipitation, potassium acetate and NP-40 were added to the lysates to 150 mM and 0.1%, respectively. Reaction mixtures were rocked at 4 °C for 2 h and beads were washed five times with washing buffer (30 mM HEPES, pH 7.3, 150 mM potassium acetate, 2 mM magnesium acetate, 5 mM DTT, 0.1% NP-40 and 1 mg ml⁻¹ Pefablock SC). After immunoprecipitation, total RNAs were isolated from the immunoprecipitates with phenol-chloroform and precipitated with ethanol. RNAs were dephosphorylated with CIP (NEB) and labelled for visualization with ³²P-γ-ATP using T4 polynucleotide kinase (Takara). Cloning of small RNAs associated with Piwi in OSCs was carried out as previously described²⁵. Deep-sequencing was performed on a GS FLX system (Roche).

Processing sequence tags. The adaptor sequences attached to the 5' and 3' ends of every sequence produced from the Roche/454 FLX system were removed. Ideally, the raw sequence belonged to one of the following canonical patterns: 5'-ATCGTCTCGGGATGAAA(N...)/TTTCATCCCGAGACGAT-3' and 3'-ATTGACCCGAGTTACAG(N...)/TTTCATCCCGAGACGAT-5', where (N...) represents an arbitrary nucleotide sequence for subsequent analyses. The adapters can have deletions at their 5'/3' ends. We used the NCBI bl2seq program (-p BLASTN -g F -W 10 -e 10 -D 1) to detect fragments of the adaptor sequences in the raw sequence and extracted 21,211 sequences belonging to the canonical/near-canonical patterns. Sequences containing ambiguous Ns were discarded.

Genome mapping and annotation. We followed a previously described procedure to map and annotate sequences²⁵. 21,089 sequences were mapped to the *Drosophila melanogaster* (dm3) genome. The annotation results are shown as a pie chart in Fig. 2a. Supplementary Fig. 2a compares the characteristics of small RNAs associated with ovarian Piwi, Aub, AGO3³ and OSC Piwi in terms of strand bias (left) and cloning frequency (right). The strand bias is computed by comparing the mapping strand of a piRNA with the strand of a coinciding transposon (112 popular classes are used) of the natural transposable element (FlyBase²⁸) track of the UCSC GenomeBrowser for Functional RNA²⁹. To distinguish a piRNA solely mapped on to a single genomic locus from piRNAs mapped on to multiple genomic loci, we use a term 'unique mapping'.

Sequence logo. We used the Web Logo program³⁰ to produce Fig. 2b and Supplementary Fig. 2c. The sequences, ranging from 25 to 27 bases, are aligned to the 5' end with appended '-' to fill the gaps for sequences shorter than 27 bases.

Frequency map. Frequency maps for OSC piRNAs were generated (Fig. 2c and Supplementary Fig. 3) representing the number of piRNAs that are 100% identical to the genomic sequence in a 5-kb sliding window. piRNAs in the forward strand are rendered in green whereas piRNAs in the reverse strand are rendered in red.

Surveying ten-base binding partners. We performed a sequence similarity search using the NCBI BLASTN program among small RNAs associated with ovarian Piwi, Aub, AGO3³ and OSC Piwi to detect potential binding partners that bind each other with 10-base full complementarity. We used the sequences that were 100% identical to the genomic sequence.

Northern blot, western blot and β-elimination analyses. Northern blot analysis was carried out essentially as previously described¹⁴. Total RNAs of OSCs and S2 cells were isolated using ISOGEN (Wako). Total RNAs from the immunoprecipitates were isolated with phenol-chloroform and precipitated with ethanol. DNA fragments for detecting *tj* transcripts (accession number AY325814) were cloned into pBS SK+. The ORF probe corresponded to nucleotides 801–1300 and the 3' UTR probe corresponded to nucleotides 2143–2402. Sequences of the primers were as follows: *tj*-ORF-forward, 5'-CATGACATGATGTGGCTGAC-3'; *tj*-ORF-reverse, 5'-GATCTGTCGAGCTGGCG-3'; *tj*-utr-forward, 5'-TTTCAAGAGAAGTGCATTCCC-3'; *tj*-utr-reverse, 5'-TATCTCATCTATCTCAATCTCTGTC-3'. PCR products were used as templates. DNA probes were synthesized using a random primed labelling kit (Takara) in the presence of ³²P-dCTP. For small RNA northern blot analyses, probes used for miR-310, *tj*-piR-1, *tj*-piR-2, *klp10a*-piR-1 and *brat*-piR-1 were as follows: miR-310, 5'-AAAGCCGGGAAGTGTGCAATA-3'; *tj*-piR-1, 5'-GGTAATGGGAATGCACCTCTCTTGAA-3'; *tj*-piR-2, 5'-TCTCATCTATCTCAATCTCTGTGCA-3'; *klp10a*-piR-1, 5'-GATGTCAGTTCGGTTTGGCGTGTGA-3'; *brat*-piR-1, 5'-TTGTGTCGCGGTTTCGGTTTGGGT-3'. The DNA oligonucleotides were labelled with T4 polynucleotide kinase in the presence of ³²P-γ-ATP. Western blot analysis was performed as described previously⁴. Ten micrograms of protein from each sample was loaded on gels. Culture supernatants of anti-Piwi hybridoma cells (P3G11)¹⁴, a mouse monoclonal antibody for Aub¹⁶, culture supernatants of anti-AGO3 hybridoma cells⁴, a rabbit polyclonal antibody to Vasa (1:1,000 dilution), a guinea-pig antibody to TJ (1:1,000 dilution)⁹ and anti-tubulin (from DSHB) (1:5,000 dilution) were used. Periodate oxidation/β-elimination treatment was performed as previously described¹⁷.

RT-PCR analyses. Total RNAs of ovaries, S2 cells and OSCs were isolated using ISOGEN according to the manufacturer's instructions. Total RNAs were treated with DNase to eliminate DNA contamination. Five-hundred nanograms of total RNA was annealed with an oligo-dT primer. Reverse transcription was performed using Accuscript High Fidelity Reverse Transcriptase (Stratagene) according to the manufacturer's instructions. The resultant cDNAs were amplified using KOD-plus DNA polymerase and primers for each gene. Sequences of the oligonucleotide primers used are: *Dmhen1*/*Pimet*, 5'-ATGTTTTCGCACAAGTTTATTTCGCGG (forward), 5'-GCCCAACACCAAGTCTTGAAC (reverse); *AGO3*, 5'-GCGAGACGAAGTACGGTCAGATAAC (forward), 5'-CAATCAAATAAGCCAATTGTGTAGCG (reverse). For quantitative RT-PCR, total RNA (0.1 µg) was used to reverse transcribe target sequences using a Transcriptor First strand cDNA Synthesis Kit (Roche) according to the manufacturer's instructions. The resulting cDNAs were amplified with a LightCycler 480 Real-Time PCR Instrument II (Roche) using the LightCycler 480 SYBR Green I Master (Roche). The primers used are shown in Supplementary Table 2.

Myc-tagged protein expression and RNAi in OSCs. A short DNA fragment encoding a Myc-tag was inserted into the KpnI site of pAc5.1/V5-HisA (Invitrogen) to produce pAcM. Full-length *piwi* and *zuc* cDNA were inserted into pAcM to yield Myc-Piwi wild type and Myc-Zuc, respectively. To produce Myc-Piwi-ΔN, a partial fragment of the *piwi* gene (amino acids 73–843) was amplified and inserted into pAcM. Mutagenesis to yield Myc-Piwi-DDAA and Myc-Piwi-ΔN-DDAA was performed with a QuikChange PCR kit using Pfu Turbo (Stratagene). Primers used are shown in Supplementary Table 2. Trypsinized OSCs (3 × 10⁶ cells) were suspended in 100 µl of Solution V of the Cell Line Nucleofector Kit V (Amaxa Biosystems) and mixed with 5 µg of expression plasmid or 200 pmol of siRNA duplex. Transfection was conducted in an electroporation cuvette using the Nucleofector instrument (Amaxa Biosystems). The transfected cells were transferred to fresh OSC medium and incubated at 26 °C. In the case of RNAi, cells were transfected again after 48 h (the detailed protocol is available on request). The siRNA duplexes used are shown in Supplementary Table 2.

Immunostaining. OSCs were fixed with 4% formaldehyde in PBS for 15 min and treated with 0.1% Triton X-100 in PBS for 15 min. After blocking with 2% BSA in PBS, they were incubated with the primary antibody for 1 h. Immunostaining of larval ovaries was performed essentially as previously described²⁶. Briefly, third larval or adult ovaries were dissected manually in PBS. The samples were fixed in 4% paraformaldehyde in PBS for 20 min and treated with methanol and detergent.

The samples were washed three times (20 min each) in PBTx (PBS containing 0.1% Triton X-100) and blocked in PBTxb (PBTx containing 5% bovine serum (Sigma)) for 1.5 h. Subsequently, samples were incubated with primary antibodies in PBTxb for 16 h at 4 °C. After washing three times (20 min each) in PBTx, samples were incubated with secondary antibodies in PBTxb for 16 h at 4 °C and rinsed three times (20 min each) in PBTx. Culture supernatants of anti-Piwi hybridoma cells (P4D2, 1:1 dilution)¹⁴, a mouse monoclonal antibody to Myc-tag (1:1,000 dilution, Sigma), a guinea-pig antibody to TJ (1:2,000 dilution)⁹, a rabbit polyclonal antibody to Vasa (1:1,000 dilution) and a mouse antibody to FasIII (1:50 dilution) were used for primary antibodies. Alexa-488-conjugated anti-mouse IgG (Molecular Probes), Cy3-conjugated anti-mouse IgG (Sigma), Alexa-546-conjugated anti-rabbit IgG (Molecular Probes), Alexa-633-conjugated anti-rabbit IgG (Molecular Probes), Alexa-546-conjugated anti-guinea-pig IgG (Molecular Probes) and Alexa-488-conjugated anti-guinea-pig IgG (Molecular Probes) were used as the secondary antibodies. DNA was stained with DAPI. All images were collected using a confocal microscope (Zeiss LSM5 EXCITER).

ChIP assay. ChIP assays were performed using the EZ ChIP Chromatin Immunoprecipitation Kit (Upstate). Briefly, 2×10^7 cells were crosslinked with

1% formaldehyde for 10 min at room temperature and sheared to produce crosslinked DNA (~200–1,000 bp in length). The DNA–protein complexes were immunoprecipitated overnight with either non-immune IgG antibody or anti-TJ antibody. After reversing crosslinking at 65 °C for 12–16 h, DNA was recovered. Short DNA fragments corresponding to *piwi* gene were amplified by a LightCycler 480 Real-Time PCR Instrument II using SYBR Premix Ex Taq (Takara). The primers used are listed in Supplementary Table 2.

27. Miyoshi, K., Tsukumo, H., Nagami, T., Siomi, H. & Siomi, M. C. Slicer function of *Drosophila* Argonautes and its involvement in RISC formation. *Genes Dev.* **19**, 2837–2848 (2005).
28. Wilson, R. J., Goodman, J. L., Strelets, V. B. & FlyBase Consortium. FlyBase: integration and improvements to query tools. *Nucleic Acids Res.* **36**, D588–D593 (2008).
29. Mituyama, T. *et al.* The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.* **37**, D89–D92 (2009).
30. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo, a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

LETTERS

Exploitation of binding energy for catalysis and design

Summer B. Thyme^{1,3}, Jordan Jarjour^{2,5}, Ryo Takeuchi⁶, James J. Havranek⁷, Justin Ashworth^{1,3}, Andrew M. Scharenberg^{2,5}, Barry L. Stoddard^{3,6} & David Baker^{1,3,4}

Enzymes use substrate-binding energy both to promote ground-state association and to stabilize the reaction transition state selectively¹. The monomeric homing endonuclease I-AniI cleaves with high sequence specificity in the centre of a 20-base-pair (bp) DNA target site, with the amino (N)-terminal domain of the enzyme making extensive binding interactions with the left (–) side of the target site and the similarly structured carboxy (C)-terminal domain interacting with the right (+) side². Here we show that, despite the approximate twofold symmetry of the enzyme–DNA complex, there is almost complete segregation of interactions responsible for substrate binding to the (–) side of the interface and interactions responsible for transition-state stabilization to the (+) side. Although single base-pair substitutions throughout the entire DNA target site reduce catalytic efficiency, mutations in the (–) DNA half-site almost exclusively increase the dissociation constant (K_D) and the Michaelis constant under single-turnover conditions (K_M^*), and those in the (+) half-site primarily decrease the turnover number (k_{cat}^*). The reduction of activity produced by mutations on the (–) side, but not mutations on the (+) side, can be suppressed by tethering the substrate to the endonuclease displayed on the surface of yeast. This dramatic asymmetry in the use of enzyme–substrate binding energy for catalysis has direct relevance to the redesign of endonucleases to cleave genomic target sites for gene therapy and other applications. Computationally redesigned enzymes that achieve new specificities on the (–) side do so by modulating K_M^* , whereas redesigns with altered specificities on the (+) side modulate k_{cat}^* . Our results illustrate how classical enzymology and modern protein design can each inform the other.

Enzymes use interactions with the substrate to promote catalysis both by bringing the substrate into close proximity and proper alignment with catalytic groups on the enzyme and by selectively stabilizing the transition state for the chemical reaction^{3–5}. Dissection of the contributions to enzyme catalysis has taken on renewed importance with the advent of computational and directed evolution approaches for engineering novel enzymatic activities for applications ranging from synthetic chemistry to therapeutics^{6,7}. Reprogramming the specificity of the LAGLIDADG family of homing endonucleases for genome engineering and biotechnology purposes is one such application^{8,9}.

Control experiments probing the binding specificity of the I-AniI homing endonuclease, in preparation for computational redesign of specificity, revealed a striking asymmetry in the effect of base substitutions on binding affinity (Fig. 1a). DNA cleavage and DNA binding by Y2 I-AniI endonuclease¹⁰ were assayed for 60 different target sites, each containing a single base-pair substitution from the wild-type recognition sequence. Consistent with previous observations¹¹, enzyme activity assays showed that many nucleotide substitutions

throughout the extended 20-bp recognition site abrogated or reduced cleavage, reflecting the high sequence specificity of the endonuclease (Fig. 1b). Fluorescence-binding experiments showed that for mutations between –10 and –3 on the (–) side of the interface, this loss of cleavage activity is associated with a loss of binding affinity. In sharp contrast, mutations in the –2 to +10 region of the recognition site, which also eliminated or reduced cleavage, had a minimal effect on substrate binding (Fig. 1c).

To determine whether the differences between the (–) and (+) side substitutions reflected differential contributions to ground-state association versus transition-state stabilization, the extent of cleavage of a linear double-stranded template as a function of time was determined for all 60 singly substituted sites under single-turnover conditions, and pseudo-Michaelis–Menten parameters¹² K_M^* and k_{cat}^* were obtained from these data (Supplementary Figs 1–3). Comparison of k_{cat}^*/K_M^* for related substrates highlights the high sequence specificity of the enzyme: for example, at position –4 k_{cat}^*/K_M^* for the wild-type G:C base pair is more than 2,000-fold greater than for A:T and more than 400-fold greater than for C:G (Fig. 1b and Supplementary Table 1; because specificity is determined by the differences in k_{cat}^*/K_M^* for different substrates¹³, these results provide perhaps the most rigorous quantification of homing endonuclease specificity so far). The contribution of target-site interactions to ground-state stabilization (K_M^* , Fig. 1d) versus transition-state stabilization (k_{cat}^* , Fig. 1e) was found to be skewed: substitutions on the (–) side increased K_M^* significantly without reducing k_{cat}^* , whereas substitutions on the (+) side decreased k_{cat}^* with little effect on K_M^* . The overall segregation of the kinetic contributions to specificity is shown graphically in Fig. 1f and in the structural schematic in Fig. 1a: most single-base substitutions in the target affect k_{cat}^* (blue, (+) side) or K_M^* (red, (–) side) but not both. The striking feature of our results is that the apparent symmetry of the binding interface is completely broken during catalysis: chemically, very similar protein–DNA contacts are used for substrate association on the left side and selective transition-state stabilization on the right side.

Surface display methods are widely used to engineer proteins with new binding specificities¹⁴. The sequence specificity profile obtained for singly substituted target sites binding to I-AniI displayed on the surface of yeast closely parallels the profile observed in the solution fluorescence experiments (Fig. 1c) (J.J., B.L.S. and A.M.S., unpublished observations). We reasoned that cleavage of mutated target sites with increased K_M^* should be suppressible by tethering the DNA duplex containing the target site adjacent to the displayed enzyme on the yeast surface; the increase in local substrate concentration should compensate for the decrease in ground-state binding affinity (Fig. 2b). Indeed, mutations between positions –10 and –3 (red) that greatly reduced binding in solution do not slow tethered cleavage on the yeast cell surface (Fig. 2c). In contrast, substitutions on the right side of the target

¹Department of Biochemistry, ²Department of Immunology, ³Graduate Program in Biomolecular Structure and Design, ⁴Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA. ⁵Seattle Children's Hospital Research Institute, 1900 9th Ave M/S C9S-7, Seattle, Washington 98177, USA. ⁶Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue, Seattle, Washington 98109, USA. ⁷Department of Genetics, Campus Box 8232, Washington University School of Medicine, 4566 Scott Avenue, St Louis, Missouri 63110, USA.

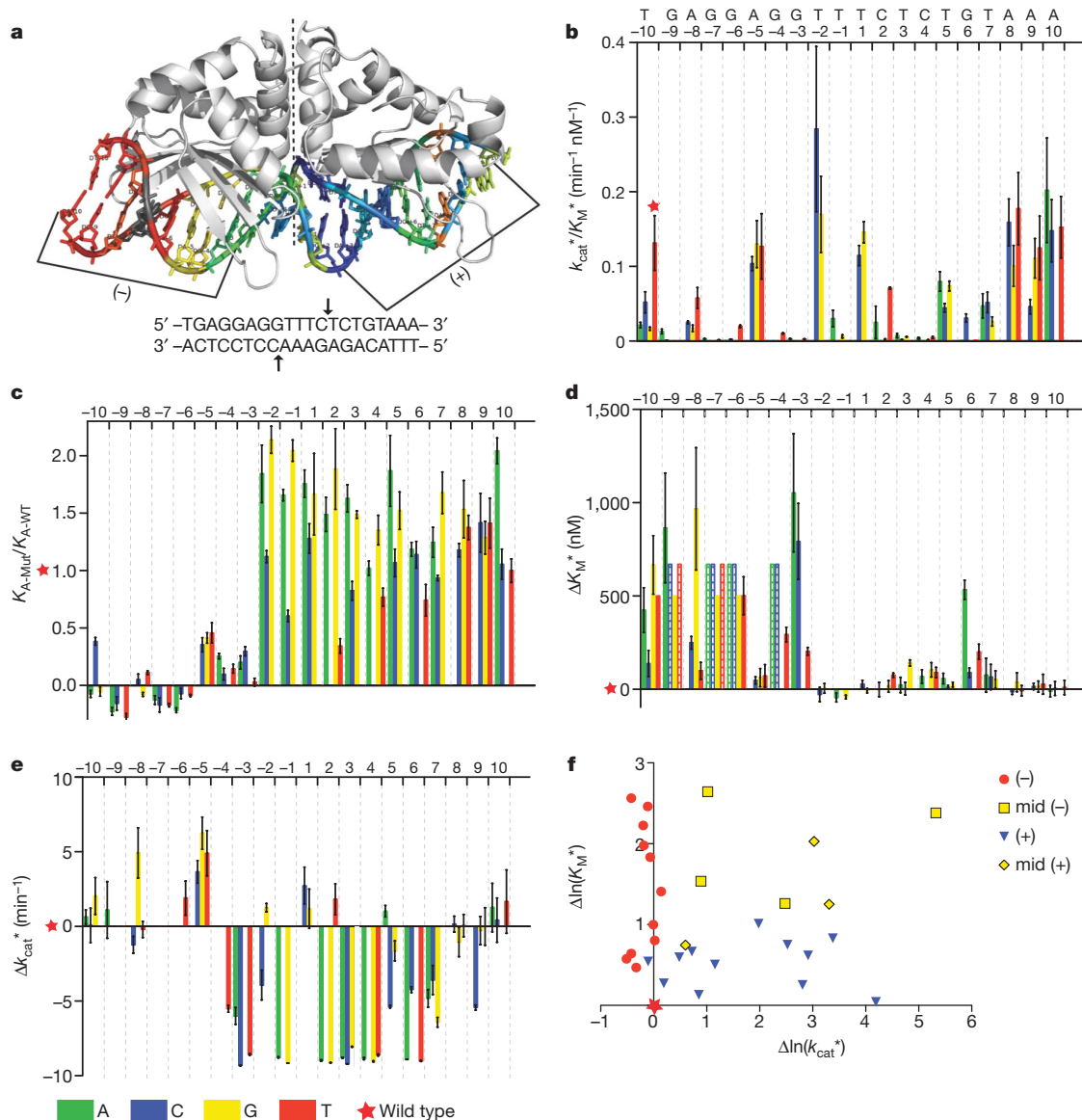


Figure 1 | Segregation of contributions to binding and catalysis. **a**, Ribbon diagram of the I-AniI enzyme in complex with the wild-type target site (2QOJ¹¹). Target site and positions of DNA cleavage are shown below: (–) side cleavage site is cut before (+) side site¹⁰. **b–e**, Colour scheme: A, green; C, blue; G, yellow; T, red; error bars, s.e.m. **b–e**, k_{cat}^*/K_M^* values for the wild-type target site (red star) and each of the 60 singly substituted target sites (vertical bars). Substitutions throughout the length of the target site abrogate enzyme activity, demonstrating the high sequence specificity of the enzyme. **c**, Relative binding affinities determined for each singly substituted target site using fluorescence competition assays. Substitutions on the left side, but not the right side, significantly reduce binding affinity. **d**, K_M^* values for each singly substituted target site relative to the wild type. As in **c**, substitutions on the left but not the right display significantly different values from wild type. **e**, k_{cat}^* values for each singly substituted target site relative to the wild-type site. In contrast to **c** and **d**, substitutions between site (blue) that reduced cleavage in solution also reduced enzyme activity in the tethered cleavage assay, consistent with their reduction of k_{cat}^* . Substitutions that disrupt interactions involved in selective transition-state stabilization cannot be overcome by increasing the local concentration of substrate.

Assuming the simple free-energy diagram in Fig. 2a, we can make inferences from the kinetic data in solution and on the yeast surface about the structures of the Michaelis and transition-state complexes. Side-chain–base-pair interactions from positions –10 to –5 are present in both the Michaelis complex and the transition state (base substitutions increase K_M^* and K_D in solution and do not affect k_{cat}^*

or the rate when tethered). Sequence-specific base-pair interactions from +3 to +8 are formed only in the transition state (substitutions have no effect on K_M^* or K_D , reduce k_{cat}^* , and slow the rate when tethered). A third class of interactions (at –5 and +7 for example) appear to be formed in the Michaelis complex but not the transition state (substitutions increase or decrease both k_{cat}^* and K_M^*/K_D).

Importantly for the design calculations described in the next section, three observations suggest that the crystal structure of the complex likely resembles the transition state more than the Michaelis complex: (1) specific interactions on the (+) side of the DNA target present in the crystal structure appear to be formed in the transition

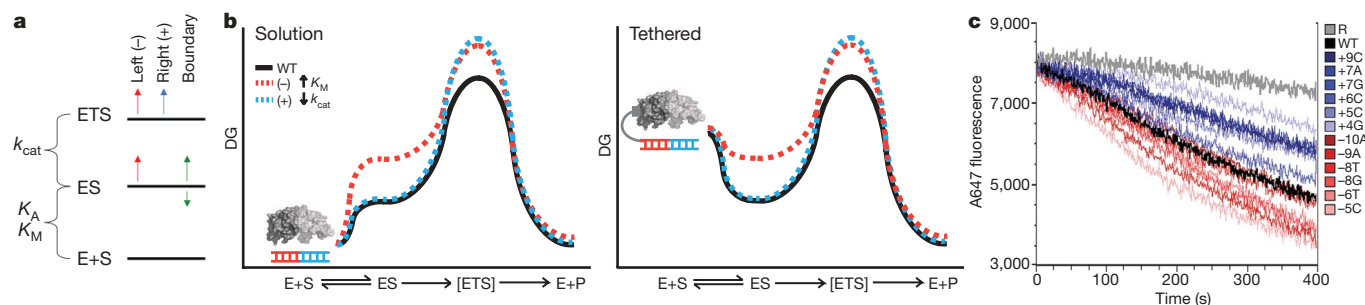


Figure 2 | Contributions to catalysis. **a**, Free-energy diagram showing the effect of target-site substitutions on the free energies of substrate binding and transition-state stabilization. ETS, enzyme–transition-state complex; ES, enzyme–substrate complex; E+S, free enzyme and substrate; E+P, free enzyme and product; DG, free energy difference. Most substitutions on the left side increase K_M^* , and hence the free energy difference between ES and E+S, but leave k_{cat}^* , and hence the free energy difference between ETS and ES, unchanged. These substitutions thus seem to disrupt interactions made in both states. Most substitutions on the right side instead leave K_M^* unchanged but decrease k_{cat}^* , and hence raise only ETS, suggesting they remove interactions present in ETS but not ES. A small subset of positions (labelled boundary) appears to stabilize or destabilize ES selectively while not affecting ETS; these substitutions may disrupt interactions present in ES but not in ETS. **b**, Free-energy profiles for a free (left) and tethered (right) system. Red profile: substitutions that remove interactions present in both ES and ETS;

state but not the Michaelis complex; (2) the third class of substitutions mentioned above that appear to stabilize only the Michaelis complex make few interactions in the crystal structure (Supplementary Fig. 4); and (3) calculations of Rosetta specificity based on the crystal structure correlate better with catalytic efficiency than with binding affinity (Supplementary Fig. 5).

Monomeric LAGLIDADG homing endonucleases, which recognize non-palindromic targets, are attractive scaffolds for genome engineering applications¹⁵. An important challenge is to reprogram the substrate specificity of these enzymes towards desired target sequences⁹. To redesign I-AniI specificity using Rosetta¹⁶, the target site in the crystal structure of the I-AniI protein–DNA complex is mutated *in silico* and the program searches for combinations of amino-acid substitutions that allow the formation of energetically favourable interactions with the new base pairs, but not with the wild-type base pairs¹⁶. Design calculations were performed for six target-site variants bearing single base-pair substitutions, genes encoding the amino-acid sequences of eight redesigned enzymes were constructed and the enzymes purified. DNA cleavage assays revealed that the designed specificity changes were for the most part achieved (Supplementary Table 2). These results demonstrate that I-AniI cleavage specificity can be reprogrammed by computational protein design, thereby providing starting points for the larger-scale specificity changes required to cleave physiological target sites.

An enzyme redesigned for a new target site could achieve altered specificity either by changing k_{cat}^* , changing K_M^* or changing both. To determine whether the designed changes in specificity were a result of changes in K_M^* or k_{cat}^* , for each of eight designed endonucleases we measured the single-turnover cleavage kinetics for target substrates containing each of the four possible base pairs at the redesign position (Supplementary Table 2). A design aimed at specific recognition of a DNA target site containing base pair –8G:C (Fig. 3a) achieved specificity exclusively by modulating K_M^* : K_M^* decreased for G:C, and increased for A:T, T:A and C:G. In contrast, a design aimed at specific recognition of +8C:G (Fig. 3b) achieved specificity entirely through k_{cat}^* : k_{cat}^* decreased for A:T, G:C and T:A, but was unchanged for +8C:G. Both of these designed enzymes have high specificity at neighbouring base pairs, and overall specificities that are higher than the wild-type enzyme in the targeted regions (Supplementary Fig. 6). A design aimed at specific recognition of the –3C:G substitution (Fig. 3c), at the boundary between

blue line, substitutions that remove interactions present in ETS but not ES. Tethering increases the free energy of free E+S to the point that the rate depends only on the free-energy difference between ES and ETS. Because this free-energy difference is unchanged by substitutions that remove interactions made in both ES and ETS (red profile), they do not affect the rate in the tethered case. **c**, Cleavage kinetics of endonuclease tethered to yeast cells.

Surface displayed enzyme cleaves a tethered fluorescently labelled oligonucleotide, which then diffuses away from the yeast surface resulting in loss of fluorescence. Black, wild-type target site; random DNA, grey; shades of red, left-side target-site substitutions; shades of blue, right-side target-site substitutions. Tethering suppresses decreases in cleavage rate produced by (–) side but not (+) side mutations. The rate increases observed for some (–) side variants such as –5c were consistent across multiple independent experiments and with the *in vitro* results in Fig. 1e.

K_M^* - and k_{cat}^* -influencing positions (Fig. 1), displayed changes in both k_{cat}^* and K_M^* , consistent with the results with the wild-type enzyme at this position. These trends hold for the remaining designs as well (Supplementary Table 2 and Supplementary Fig. 7): we find generally that the left-side designs achieve specificity primarily by modulating ground-state binding affinity, whereas the right-side designs achieve specificity by modulating the stability of the transition state.

Our results suggest that initial binding of I-AniI to its target site involves formation of base-specific interactions on the (–) side and lower-affinity non-specific interactions on the (+) side to form the Michaelis complex (the latter are suggested by yeast display experiments which show that the enzyme binds less tightly to the (–) half-site than to the full site (J.J., B.L.S. and A.M.S., unpublished observations)). Catalysis then requires bending of the DNA (note bend in Fig. 1a), which is stabilized at the transition state by newly formed specific interactions between the (+) side and the enzyme. Such a two-stage mechanism (see Supplementary Information section C) may be a general solution to the problem of specific target-site recognition by enzymes that act on distorted DNA substrates. If the enzyme only bound to the distorted site, binding would require enzyme to be at the site (which may occur only once in the genome) simultaneous with fluctuation of the DNA into the distorted conformation; because both are rare events, the net rate of binding, the product of two small numbers, would be very slow. If, instead, the enzyme can bind with some sequence specificity to undistorted target sites, the probability of being close enough to capture (and perhaps promote) fluctuations that distort the DNA will be very much higher. In I-AniI the total transition-state binding energy appears to be roughly divided between the two steps: the N-terminal domain guides the enzyme to potential target sites that match on the (–) side, and the C-terminal domain specifically stabilizes the transition state if there is also a match on the (+) side.

There is considerable synergy between classical enzymology and modern computational design. Design should be informed by detailed analyses of the wild-type enzyme because, depending on the enzyme and substrate concentrations in the application the designed enzymes are to be used for, it may be necessary to re-engineer K_M , k_{cat} and/or k_{cat}/K_M . Conversely, computational design can provide insight into the basis for transition-state stabilization. Our kinetic dissection of I-AniI cleavage activity also has implications for endonuclease

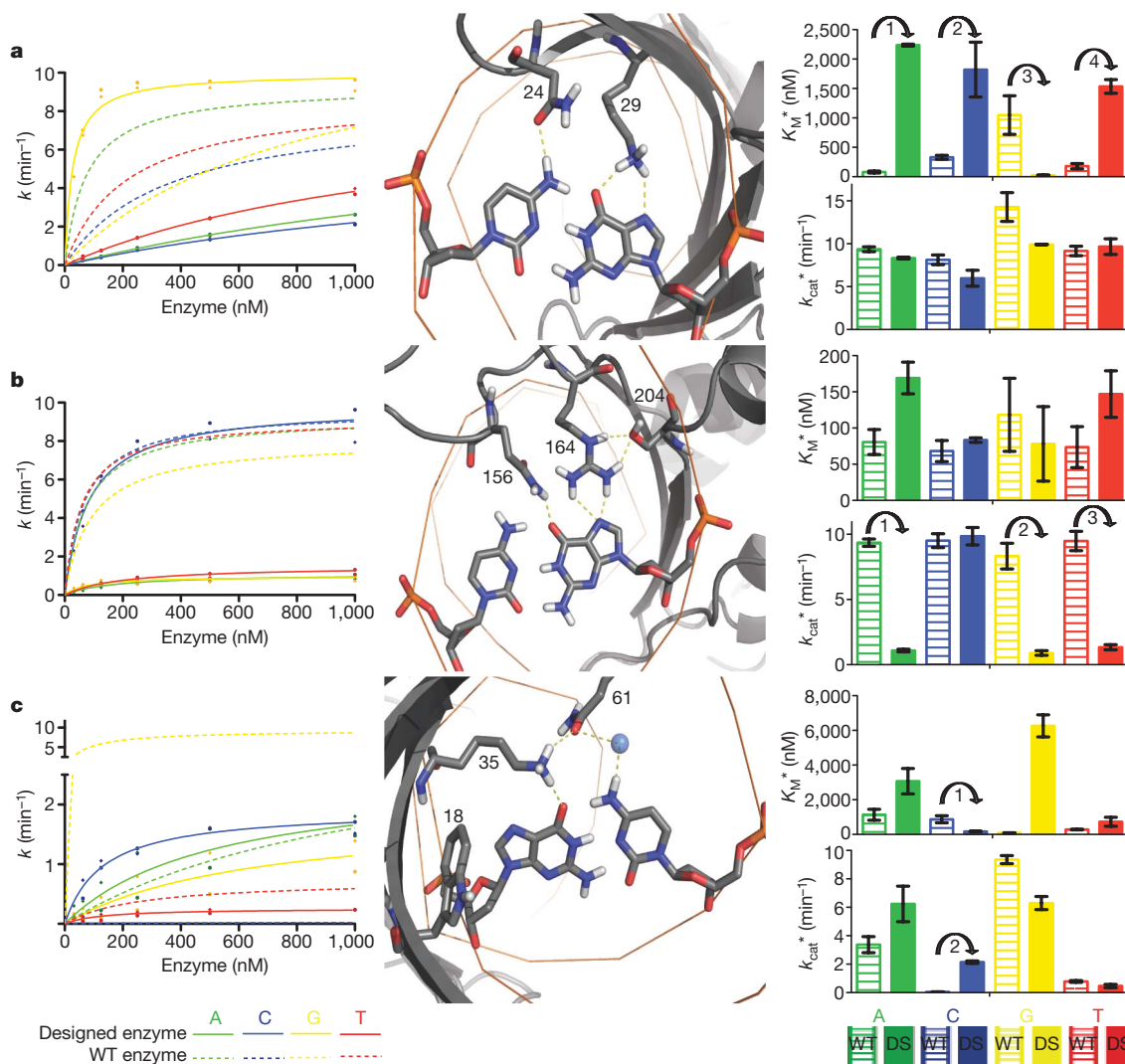


Figure 3 | Computational redesign of specificity. Colour scheme: A, green; C, blue; G, yellow; T, red; error bars in right panels, s.e.m. **a**, Design for -8A:T to -8G:C substitution (K24N, T29K). Middle panel: the designed residues N24 and K29 make direct hydrogen bonds to -8G and -8C , respectively. Left panel: the concentration dependence of the cleavage activity for the designed enzyme (solid lines) for different base pairs at the -8 position differs considerably from the wild-type enzyme (dashed lines). Right panel: the k_{cat}^* values remain approximately the same for both the wild-type and designed enzymes against all target sites, but the K_{M}^* values are decreased for the target G base pair (arrow 3) and increased significantly for the other three substitutions (arrows 1, 2 and 4). **b**, Design for $+8\text{A:T}$ to $+8\text{C:G}$ substitution (L156Q, I164R, T204S). Middle panel: designed residues R164 and Q156 make direct hydrogen bonds to $+8\text{G}$. Designed

residue S204 holds R164 in position. The kinetic traces (left panel) and bar graphs (right panel) show this design achieves altered specificity through changing k_{cat}^* . The K_{M}^* values remain approximately the same for both the wild-type and designed enzymes against all target sites, but the k_{cat}^* values are significantly decreased for all of the competitor target sites (arrows 1, 2 and 3). **c**, Design for -3G:C to -3C:G substitution (Y18W, E35K, R61Q). Middle panel: designed residues K35 and Q61 make a direct hydrogen bond to -3G and a water-mediated hydrogen bond to -3C , respectively. Q61 and K35 also hydrogen bond with each other, and designed residue W18 further helps position K35 through packing interactions. The kinetic traces (left panel) and bar graphs (right panel) show this design achieves altered specificity through changing both k_{cat}^* and K_{M}^* . The designed enzyme has an increased k_{cat}^* (arrow 2) and decreased K_{M}^* for the -3C (arrow 1).

re-engineering using yeast display: selection based on binding may be sub-optimal because substrate binding could be optimized at the expense of transition-state stabilization, whereas selection for cleavage in the tethered substrate system could yield variants with decreased solution cleavage due to increased K_{M}^* . These pitfalls could potentially be overcome by selecting both for k_{cat}^* and K_{M}^* , perhaps by alternating between the two selection procedures. More generally, the union of classical enzymology with modern computational design and selection technology, as illustrated here, provides a powerful approach to revealing the mechanistic basis for, and subsequently reprogramming of, sequence-dependent molecular recognition.

METHODS SUMMARY

Experimental preparation and kinetic analysis. I-AniI was expressed in *Escherichia coli* BL21 (DE3) using a standard auto-induction protocol and purified over a His-trap column. Linearized plasmid substrates were prepared for

each of the 60 singly substituted target sites. Kinetic assays were performed over a 20-fold range of enzyme concentrations (from 30 to 1,500 nM, depending on the substrate) with 5 nM DNA substrate, and analysed by agarose gel electrophoresis followed by integration of product and substrate band densities. The velocity versus enzyme concentration profiles were determined two to four independent times; reported k_{cat}^* and K_{M}^* values are the average of those determined from the independent experiments. Fluorescence competition-binding assays were performed as previously described¹⁷.

Computational design. New target sequences were mapped on to the I-AniI-DNA crystal structure (2QOJ¹¹) and the Rosetta computational design methodology was used to optimize the amino-acid sequence of the protein to maximize affinity for the new site¹⁶. The predicted specificity of the resulting protein models for the desired target sequence was computed using Rosetta, and designs that were predicted to bind tightly and specifically were subjected to further optimization using flexible backbone protein design (Supplementary Methods). The tightest binding and most specific designs were again selected, and the designed amino-acid substitutions were removed one at a time. If no

significant loss was predicted in either specificity or binding energy, the substitution was removed from the design. The '-8G:C_A' (K24N/T29K) and '-8G:C_B' (K24N/T29Q) designs were generated instead using a genetic algorithm to optimize binding affinity and specificity simultaneously (Supplementary Methods). Genes encoding the designed proteins were assembled from oligonucleotides, and the designed proteins were expressed, purified and assayed as described above.

Tethered cleavage on yeast surface. PCR-generated DNA substrates, labelled with biotin and Alexa 647, were tethered by an antibody–streptavidin–phycoerythrin bridge to the haemagglutinin epitope of I-Anil expressed on the surface of *S. cerevisiae* in conditions that prohibited catalysis. Samples were then spiked with 10 mM MgCl₂ and placed in a pre-warmed chamber at 37 °C and fluorescence measurements were acquired on a flow cytometer. The Alexa 647 signal from a phycoerythrin-normalized population of each sample was then plotted against time to generate the curves shown.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 June; accepted 15 September 2009.

- Jencks, W. P. Mechanism of enzyme action. *Annu. Rev. Biochem.* **32**, 639–676 (1963).
- Bolduc, J. M. *et al.* Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.* **17**, 2875–2888 (2003).
- Wells, T. N., & Fersht, A. R. Use of binding energy in catalysis measured by mutagenesis of tyrosyl-tRNA synthetase. *Biochemistry* **25**, 1881–1886 (1986).
- Fersht, A. R. Relationships between apparent binding energies measured in site-directed mutagenesis experiments and energetics of binding and catalysis. *Biochemistry* **27**, 1577–1580 (1988).
- Benkovic, S. J. & Hammes-Schiffer, S. A perspective on enzyme catalysis. *Science* **301**, 1196–1202 (2003).
- Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
- Collins, C. H., Yokobayashi, Y., Umeno, D. & Arnold, F. H. Engineering proteins that bind, move, make, and break DNA. *Curr. Opin. Biotechnol.* **14**, 371–378 (2003).
- Smith, J. *et al.* A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Res.* **34**, e149 (2006).
- Redondo, P. *et al.* Molecular basis of xeroderma pigmentosum group C DNA recognition by engineered meganucleases. *Nature* **456**, 107–111 (2008).
- Takeuchi, R., Certo, M., Caprara, M. G., Scharenberg, A. M. & Stoddard, B. L. Optimization of *in vivo* activity of a bifunctional homing endonuclease and maturase reverses evolutionary degradation. *Nucleic Acids Res.* **37**, 877–890 (2008).
- Scalley-Kim, M., McConnell-Smith, A. & Stoddard, B. L. Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.* **372**, 1305–1319 (2007).
- Halford, S. E., Johnson, N. P. & Grinstead, J. The EcoRI restriction endonuclease with bacteriophage lambda DNA. Kinetic studies. *Biochem. J.* **191**, 581–592 (1980).
- Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Analysis and Protein Folding* (W. H. Freeman, 1998).
- Gai, S. A. & Wittup K. D.. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17**, 467–473 (2007).
- Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature Biotechnol.* **26**, 808–816 (2008).
- Ashworth, J. *et al.* Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **441**, 656–659 (2006).
- Zhao, L., Pellenz, S. & Stoddard, B. L. Activity and specificity of the bacterial PD-(D/E)XK homing endonuclease I-Ssp6803I. *J. Mol. Biol.* **385**, 1498–1510 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by a National Science Foundation graduate research fellowship to S.B.T., the US National Institutes of Health (GM084433 and RL1CA133832), the Foundation for the National Institutes of Health through the Gates Foundation Grand Challenges in Global Health Initiative, and the Howard Hughes Medical Institute. We thank A. Quadri for help with plasmid substrate preparation and M. Scalley-Kim for I-Anil cleavage data collected in the presence of Mn²⁺.

Author Contributions S.B.T. and J.J.H. performed computational design calculations and S.B.T. performed kinetic characterization of all designed and wild-type enzymes. R.T. performed the fluorescence competition binding experiment. J.J. performed the surface-expressed tethered cleavage assay. J.A. and J.J.H. developed computational design procedures. S.B.T. and D.B. wrote the paper. Multiple discussions of shared data among all authors at Northwest Genome Engineering Consortium (http://research.seattlechildrens.org/centers/immunity_vaccines/ngec/) group meetings contributed to the recognition of binding/catalysis asymmetry in I-Anil Y2 and the conceptual development of this manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.B.T. (sthyme@u.washington.edu) or D.B. (dabaker@u.washington.edu).

METHODS

Protein expression and purification. Genes encoding Y2 I-Anil^{10,18} designs were assembled from oligonucleotides¹⁹, cloned into a variant of the pet15 expression vector and sequence-verified plasmids were transformed into BL21 Star (Invitrogen). A 1-l culture of auto-induction media²⁰ was inoculated with several colonies, grown at 37 °C for about 12 h (to approximately saturation) and expression at 18 °C continued for about 24 h. Cells were harvested, resuspended in Tris 20 mM pH 7.5, 1.0 M NaCl and 30 mM imidazole, and lysed by sonication and lysozyme. The soluble fraction was loaded onto a 1 mL HisTrap FF crude column (GE Healthcare) and I-Anil variants were purified by imidazole gradient elution on an AKTA express (GE Healthcare). The proteins were concentrated and the buffer was exchanged to Tris 20 mM pH 7.5, 500 mM NaCl and 50% (v/v) glycerol for storage. Purity of the proteins was assessed by SDS–polyacrylamide gel electrophoresis and the concentration of samples with about >95% purity was determined by measuring the absorbance at 280 nm using the calculated extinction coefficient²¹. The concentration of enzyme in the <95% pure samples was determined by generating a standard curve with a pure I-Anil protein, correlating protein concentration with band density (calculated with ImageJ (<http://rsbweb.nih.gov/ij/>)) and comparing the band density of the I-Anil protein in impure samples run on the same gel as the standard curve.

Plasmid substrate construction. All single base-pair variants from the wild-type target site in pBluescript were individually constructed by site-directed mutagenesis as described²². Sequence-verified plasmids were linearized with ScaI before the kinetic assays to facilitate product identification.

Endonuclease activity assays. Kinetic assays. Previous work²³ has confirmed that I-Anil, like other LAGLIDADG endonucleases, is a single-turnover enzyme, and the conditions for single-turnover kinetics¹² were met in all experiments. The ionic strength of the enzyme reaction buffer was optimized for enzyme activity and stability to a final solution of 170 mM KCl, 10 mM MgCl₂ and 20 mM Tris, pH 9.0. Enzyme was diluted in 1.25× reaction buffer to working concentrations, serial twofold dilutions were made, and both substrate plasmid and diluted enzyme were incubated separately at 37 °C for 1 min. The appropriate amount of plasmid (one-fifth of the reaction volume) was added to each reaction for a final 1× reaction buffer and final plasmid concentration of about 5 nM (the lowest concentration still readily visible on agarose gel). The plasmid (one-fifth of reaction volume) was added to the enzyme (four-fifths of reaction volume) to minimize heat loss during the transfer (which was found to add significant noise to the data). Reactions were halted with 200 mM EDTA, 30% glycerol and bromophenol blue. DNA fragments were separated on 1.2% agarose TBE gels, which were then stained in a standard ethidium bromide solution and subsequently destained in water for maximum contrast between DNA and background. All data were collected by integrating the density of the substrate (2,959 bp) and product bands (1,801 bp and 1,158 bp) using ImageJ (<http://rsbweb.nih.gov/ij/>). The percentage of product formed is equal to the sum of the density of the two product bands divided by the total sum of the densities of the three bands. The progress curves fit to single exponentials for all enzyme concentrations (Supplementary Fig. 1) and for all target sites except for several substitutions in the central four base pairs between the cleavage sites on the two DNA strands (Supplementary Fig. 2).

Assays for specificity positions adjacent to designed nucleotide. Twofold serial dilutions of enzyme from 1500 to 11 nM were made in 1.25× reaction buffer and the enzyme was reacted with about 5 nM substrate (in 1× reaction buffer) for half an hour at 37 °C. Reactions were halted and data were analysed as described in the 'Kinetic assays' section.

Fluorescence competition binding assay¹⁷. Unlabelled DNA oligonucleotides with each of the 60 single base-pair substitutions in the I-Anil target site (wild-type I-Anil site, 5'-TGAGGAGGTTTCTCTGTAAG-3'), a negative control sequence (5'-CTCTTCTGCATATATCTCC-3'), an unlabelled wild-type site oligonucleotide and a wild-type site oligonucleotide labelled with 5' Cy3 were synthesized with six consecutive 'A' flanking on each end (Integrated DNA Technology, 100-nmol scale, salt-free). Complementary oligonucleotides were ordered for all 63 sites and double-stranded target DNA was prepared by annealing equal amounts of complementary strands.

His-tagged I-Anil was immobilized by incubating 200 µl of 100 nM I-Anil in TBS/BSA buffer (50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.2% BSA) in wells of Nickel-NTA coated HisSorb Plates (Qiagen) for 2 h at room temperature. Unbound protein was removed and the plates were washed four times with TBS/Tween-20 (50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.05% Tween-20). The immobilized I-Anil in the microtitre plate was incubated for about 4 h with both 100 nM labelled target DNA duplex and 3 µM (30-fold excess) of one unlabelled duplex per well in 200 µl of binding buffer (50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 0.02 mg ml⁻¹ poly(dI-dC), 10 mM CaCl₂). The plates were washed four times with TBS (50 mM Tris-HCl (pH 7.5), 150 mM NaCl), and the fluorescent

signal retained in each well was quantified using a SpectraMax M5/M5e Microplate Reader (Molecular Devices) (excitation, 510 nm; emission, 565 nm; cutoff, 550 nm). Additional negative control experiments performed in the absence of the enzyme indicated that no significant detectable fluorescent signal was retained after the protocol described above was completed. Relative binding affinities were calculated using the following equation: relative binding affinity = $[(F_n - F_x) \times F_i] / [(F_n - F_i) \times F_x]$, where F_x , F_i and F_n indicate fluorescent intensities obtained from wells in which the immobilized protein was incubated with the unlabelled singly substituted target sites, wild-type target site and negative control sequence, respectively.

Cleavage kinetics of endonuclease tethered to yeast cells. Surface display of I-Anil on *S. cerevisiae* was performed using standard methods¹⁴. For each sample, 5×10^6 cells were stained with biotinylated anti-HA11, followed by secondary staining with streptavidin-DNA substrate conjugates (1:3 molar ratio) on ice and in the absence of divalent cations. DNA substrates were generated by PCR using biotinylated and Alexa 647-conjugated primers complementary to the 5' and 3' sequences flanking the I-Anil target (or indicated target-site variants). PCR products were purified by ExoI digestion followed by size exclusion chromatography on a G-100 column (GE Healthcare) before conjugation with streptavidin-phycoerythrin. Samples were then spiked with 10 mM MgCl₂ and placed in a pre-warmed 37 °C chamber and acquired at an approximate event rate of $3,000 \text{ s}^{-1}$ for 400 s on a BD FACSaria II flow cytometer. Processing was performed using FloJo software (Treestar). Live cells were gated by forward- and side-scatter properties, and doublets and clumped cells were excluded on the basis of forward scatter area versus height linearity. The Alexa 647 signal from a phycoerythrin-normalized population of each sample was then plotted against time to generate the curves shown.

Computational design methods. Single-state design (designs -9C, -8G_C, -6C, +5C and +8C). The computational design of homing endonuclease-DNA specificity was performed using the Rosetta design software in a manner specifically designed to predict new protein sequences that would bind with high affinity to novel DNA sequences²⁴. The prediction of designed proteins with novel interactions to substituted base pairs in the I-Anil recognition sequence was performed by mutation and Monte Carlo repacking of amino-acid side chains as described in Ashworth *et al.* 2006 (ref. 16). The template for the design calculations was the crystal structure of the I-Anil-DNA complex (Protein Data Bank accession number 2QOJ¹¹). Additionally, minor shifts of the protein backbone were modelled only in the vicinity of the designed region using a loop-rebuilding algorithm^{25,26}. The specificity of each hypothetical new protein sequence for the intended new DNA recognition sequence was calculated as the Boltzmann probability of the intended complex versus a partition function consisting of each base-pair possibility at the redesigned DNA base pair²⁷. After design, predicted protein sequences with the most favourable binding energy and highest predicted specificity were reverted position by position to the wild-type amino-acid sequence to identify (and revert) designed mutations that did not significantly contribute to the energy or specificity of the designed complex.

Multi-state design (designs -8G_A and -8G_B). Two base-pair positions in the structure were computationally mutated to generate a partial match to a recognition site in the *IL-2R γ* gene in a mouse model of severe combined immunodeficiency disease. Specifically, positions -9G:C and -8A:T were modelled as -9A:T and -8G:C. A multistate design calculation²⁸ was performed to select amino acids at positions 24Z, 26Z, 27Z and 29Z. Three states were included in the design. The first state was the target state, which was modelled using the altered DNA structure. The second state was the original structure with the wild-type DNA sequence, which served as a competitor to enforce binding specificity of the selected proteins for the altered recognition site (negative design state). The third state was the modelled structure of the best single-state design for the target state with the modified DNA sequence; the energy associated with this state is a constant during the multi-state design procedure. It represents the best scoring protein-altered DNA complex as assessed with the Rosetta energy potential, and it is therefore impossible for the energy associated with the target state to be lower than this value. As a result, multiple calculations were performed which differed from each other only in an artificial offset applied to the third state. Progressively larger offsets bias the calculations to select sequences that achieve higher specificity for the first state over the second state at the expense of achieving Rosetta scores that are allowed to be progressively worse than the third state.

A genetic algorithm was used to evolve a population of sequences that prefer the target state to the two competitors. An initial population of 2,000 sequences was generated by selecting random amino acids at the four design positions. The side-chain conformations of these four residues (with the rest of the protein and DNA structure held fixed) were predicted for the first and second states using a Monte Carlo algorithm, and the Rosetta score recorded. As noted above, the energy of the third state is a constant. A 'fitness' score for each sequence *i* in the population is calculated: $\text{fitness}_i = E_{\text{target}} - \langle E_{\text{competitors}} \rangle$, where E_{target} is the energy of the

target state and the angled brackets denote an ensemble (Boltzmann-weighted) average over the energies of the competitors. Conceptually, the fitness corresponds to the transfer free energy of the protein from the ensemble of competitors to the target state. Subsequent generations were constructed using the following procedure. First, the sequence with the best (lowest) fitness was promoted automatically. Next, 1,980 sequences were created by recombining two members of the population using uniform crossover of two parents chosen by tournament selection.²⁹ Finally, the remaining 19 sequences were generated by mutating a single parent chosen by tournament selection with a 25% chance of randomizing each position in turn. A fitness value was calculated for each new sequence, and the population was propagated for 30 generations.

18. Bolduc, J. M. *et al.* Structural and biochemical analyses of DNA and RNA binding by a bifunctional homing endonuclease and group I intron splicing factor. *Genes Dev.* **17**, 2875–2888 (2003).
19. Stemmer, W. P. C., Cramer, A., Ha, K. D., Brennan, T. M. & Heyneker, H. L. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**, 49–53 (1995).
20. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
21. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423 (1995).
22. Kunkel, T. A., Roberts, J. D. & Zakour, R. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Methods Enzymol.* **154**, 367–382 (1987).
23. Geese, W. J., Kwon, Y. K. & Waring, R. B. *In vitro* analysis of the relationship between endonuclease and maturase activities in the bi-functional group I intron-encoded protein, I-Anil. *Eur. J. Biochem.* **270**, 1543–1554 (2003).
24. Havranek, J. J., Duarte, C. M. & Baker, D. A simple physical model for the prediction and design of protein–DNA interactions. *J. Mol. Biol.* **344**, 59–70 (2004).
25. Canutescu, A. A. & Dunbrack, R. L. Jr. Cyclic coordinate descent: robotic algorithm for protein loop closure. *Protein Sci.* **12**, 963–972 (2003).
26. Das, R. *et al.* Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69**, 118–128 (2007).
27. Ashworth, J. & Baker, D. Assessment of optimization of affinity and specificity at protein–DNA interfaces. *Nucleic Acids Res.* **37**, e73 (2009).
28. Havranek, J. J. & Harbury, P. B. Automated design of specificity in molecular recognition. *Nature Struct. Biol.* **10**, 45–52 (2003).
29. Mitchell, M. *An Introduction to Genetic Algorithms* (MIT Press, 1996).

Q&A

Ian Anderson, an ecologist at the University of Western Sydney, Australia, has won the first annual ProSPER.NET-Scopus Young Scientist award for agriculture and natural resources.



Do you think this award will change your career?

Yes, I expect it will have an impact on my career. Awards such as this can be used as a springboard for applications for funding, promotions or new job opportunities. In particular, the ProSPER.NET-Scopus award is for scientists younger than 40, which is extremely important given that science is such a difficult game in which to get permanent positions.

What is the biggest challenge for today's young scientists?

It is simply getting out of the postdoc cycle and into a permanent faculty or full-time position. There's not an excess of funding in any country, so it is very competitive to get funding and there are limited opportunities for permanent employment.

Do you have a research strategy to further your career?

My tactic up to this point has been to balance some risky, technically

challenging experiments that push the boundaries with more straightforward research. I consider the development of molecular-biology techniques, such as DNA fingerprinting of soil microbes, to be my greatest scientific achievement. It has opened up new avenues of research, helping to construct a picture of species distributions underground. But I balance that work with other experiments in which I have a more direct idea of where they will lead so that I can make the currency — publications and grant income — necessary for a young scientist to get the elusive faculty position.

Has 'sustainability' driven your research career?

In the early days of my career, I was purely interested in the ecology of a system or organism. Having studied these natural forest ecosystems for the past few years, I began to see the broader context of how organisms fit into ecosystems and into the planet in general. So,

although I may not have realized it straight away, my work on soil fungal functions in nutrient and carbon cycling has sharpened my focus on sustainability of ecosystems from a biodiversity and conservation point of view.

Is there a single question that could sustain your research career for years?

For the next 10 years, the big challenge for me will be turning my research focus to climate change. I want to understand how these soil microorganisms are going to respond to climate change and what their role is in adaptation to and potential mitigation of climate change. We are really just starting to understand how soil fungi work, so there is a huge amount of fundamental ecology to do on the organisms themselves. At the broadest level, we need to understand the basic biology of organisms in order to address bigger issues such as climate change. ■

Interview by Virginia Gewin

POSTDOC JOURNAL

The many hats of science



On my first day of graduate school, excitement and anticipation consumed me. I have to admit — I was naive. I thought learning how to conduct research effectively would be the only necessary scientific training. My experiments monopolized my thoughts, and my emotions often found themselves intertwined with the pending results. I remember going over protocols in my head while lying in bed at night to make sure that crucial steps had not been overlooked.

However, I have learned that a career in science is

much more than conducting experiments. Scientists wear many hats. They design controlled experiments, but they also assume the role of public speaker, writer, manager and mentor.

As I consider how my time as a graduate student, and now as a postdoc, has prepared me for a future in science, I think about my ability to wear these many hats. Over the years I have given many presentations. I've come to enjoy writing manuscripts and editorial articles. And I am working on improving my managerial and mentoring

skills. In overseeing a project in the lab to map the locations of hundreds of transcription factors in the yeast genome, for example, I've learned how to effectively guide others who are dealing with experimental roadblocks.

If skills are hats that scientists wear, then I've got to collect them all. It's part of being a scientist — and it gives you an edge in an increasingly competitive job market. ■

Bryan Venters is a postdoctoral fellow at the Center for Eukaryotic Gene Regulation at Pennsylvania State University, University Park.

IN BRIEF

Women honoured

This year's L'Oréal-UNESCO Awards for Women in Science have gone to five candidates. Two won for their work on toxins: microbiologist Alejandra Bravo of the National Autonomous University of Mexico in Mexico City and biochemist Lourdes Cruz of the University of the Philippines Diliman in Quezon City. Two awards were given for work on diseases: to zoologist Rashika El Ridi of Cairo University and cell biologist Elaine Fuchs of the Rockefeller University in New York. The fifth award went to Anne Dejean-Assémat, a molecular biologist at the Pasteur Institute in Paris, for her work on leukaemia and liver cancers. Each laureate will receive US\$100,000.

Geoscientist shortfall

A report from the US National Science Foundation (NSF) predicts that there will be a growing national need for geoscientists to address the problems of climate change, resource depletion, energy sustainability and environmental degradation. *GEO Vision: Unraveling Earth's Complexities through the Geosciences*, released on 15 October by the NSF's Advisory Committee for Geosciences, says geoscientists will increasingly be called on to assess how human behaviour is affecting Earth and its systems. Tim Killeen, NSF assistant director for geosciences, predicts that the 4,000 US geoscientists who graduate each year will not be enough to supply these needs.

Burnham expands south

The Burnham Institute for Medical Research in La Jolla, California, has opened a new facility in Orlando, Florida. The centre, called Burnham at Lake Nona, will employ some 250 research scientists focusing on cardiovascular pathobiology and metabolic signalling and disease. Scientific director Daniel Kelly says that the facility is recruiting 30 principal investigators to lead teams of three to five researchers and postdocs. During the 8 October dedication of the centre, the University of Florida announced that it will build a research facility and drug-development centre at the new life-sciences complex. This centre will hire 15–20 research scientists over the next three years, says Sobha Jaishankar, assistant vice-president for research at the university.



Fertile grounds

Can Brazil use its booming economy and abundant natural resources to become a life-sciences juggernaut? **Gene Russo** finds out.

Just north of central Rio de Janeiro, perched on a hilltop overlooking a busy road, is an ornate castle that seems out of place among the chaotic traffic and modest houses of the Manguinhos district. But this building (pictured above) and its surrounding 800,000-square-metre campus has had a key role in Brazil's growing science enterprise.

The Alhambra-inspired pavilion, built between 1905 and 1918, is the headquarters of the Oswaldo Cruz Foundation, commonly known as Fiocruz. Spearheaded by the Pasteur Institute-trained bacteriologist after whom it is named, the foundation helped develop vaccines to treat devastating outbreaks of bubonic plague, smallpox and yellow fever in Brazil. Cruz wanted a temple for science, an emblem people would recognize and respect. Funded by the federal government, Fiocruz is now not only central to Brazilian public health, it is also a major employer and trainer, with nearly 6,000 researchers in Rio and several other satellite campuses around the country.

Scientists and policy-makers hope that modern Brazil's temples of science will be its universities, government institutes and perhaps even its biotechnology companies.

At times, Brazilian science has been hampered by a brain drain and modest private-sector endeavours — largely the result of limited venture capital and a culture that tends not to encourage entrepreneurs. But it could become a major player in international research if it can capitalize on its vast natural

resources and booming economy.

That promise comes in part from growing government support. A decade ago, science funding was erratic, says Luiz Antônio Barreto de Castro, the secretary of R&D Policies and Programmes at Brazil's Ministry of Science and Technology. "We didn't know what the budget was going to be the next year," he says. But by 2000, the congress had approved funds for future budgets, making it possible for science-funding agencies to plan long-term projects. Taxes on oil companies and other industries provided a steady source of support. Oil giant Petrobras, for example, must by law devote 1% of its revenue towards research and development. Although the situation is greatly improved, Barreto de Castro notes that federal R&D spending is still only slightly above 1% of the country's gross domestic product. A government plan to increase that to 1.5% by 2010 is off track because of the recent financial downturn, he says.

State support

Fortunately for scientists, government money comes from the states as well. The São Paulo state government is by far the biggest contributor, sending 1% of all its tax revenue to science research — some US\$400 million each year. This investment has helped the state economy to thrive, says Barreto de Castro. Some of the other 25 states are now attempting to follow suit.

Still, for some, salaries and benefits are a concern. Hugo Armelin, a cell-signalling specialist at the state-funded Butantan

Institute in São Paulo, laments the plight of his postdocs, saying that they tend to receive moderate pay but no benefits. "In Brazil, it's easy to get fellowships for several years," he says, "but without any benefits." Postdocs, depending on their funding source, typically earn \$24,000–30,000 per year. But staff scientists at Butantan Institute, which, like Fiocruz, focuses on vaccines and anti-venoms, earn only \$20,500 per year when they start.

That can contribute to brain drain. Some fledgling Brazilian researchers who go abroad to do research or a postdoc choose to stay away because of more lucrative pay packages or more abundant opportunities. Fiocruz, for example, employs thousands of scientists, but often has 10 or 20 times more applications than it has vacancies.

For years the Brazilian government has attempted a difficult balancing act: educating scientists and forging ties at more developed research institutions abroad, while trying to ensure that the bulk of the country's science talent does not leave Brazil for good. Traditionally, the government has provided ample support for graduate study abroad. But today, such scholarships are more selective and generally target top universities, according to Barreto de Castro.

To make domestic graduate education alternatives more attractive, especially in less developed regions, Brazil's Ministry of Science and Technology has started a 'northeastern biotechnology network' programme meant to attract expertise to 30 institutions and ensure that São Paulo,

the country's richest state, is not the only destination for science talent. Most of Brazil's academic and private-sector life sciences research is concentrated in the richer areas such as São Paulo (see map).

Started in 2004, the programme has received \$15 million in the past five years. It allows students to earn their degree after stints at multiple institutions. Four hundred students receive needs-based funding and scholarships from federal and state governments. Each student's thesis draws on the research of multiple groups. In principle, the poorest states benefit most as they often don't have the money to fund biotechnology PhD programmes. The government continues to evaluate the programme's effectiveness says Barreto de Castro.

Happy returns

Many scientists who have returned to Brazil after an education or postdoc abroad have thrived. Armelin spent three years as a postdoc at the University of California, San Diego, before attending Harvard Medical School on a Guggenheim fellowship in 1982. But he decided to come back. "I realized I could be a nobody there or a somebody here," he says, laughing.

Mayana Zatz, now director of the human genome research centre at the University of São Paulo, did a postdoc at the University of California, Los Angeles. "I could have stayed," she says. "I'm glad I didn't." Zatz has had a hand in building a thriving human genetics community at São Paulo, and has probably had a much greater impact than she would have had in the United States. However, she says, there were drawbacks. Although funding has not been a problem, reagents, for example, are costly and slow to arrive, a challenge for a fast-moving, competitive field such as genetics.

The Brazilian government is also attempting to help spark private-sector science, although with mixed success. One federally funded programme offers money directly to fledgling businesses, including

biotech firms. So far, the government has put up \$200 million per year since 2006. Gerardo Mendoza, head of Bionext, a small biotech firm in São Paulo, says that this money, so far approximately \$5.9 million for his firm, has been key to keeping his business afloat. Bionext develops biocellulose for potential use in surgical repair of the brain and heart. Most of Brazil's life-science companies focus on human health and agriculture (see graphic). Money might be available, says Mendoza, but the real problem is navigating the government bureaucracy to get approval for clinical trials. "It takes a very long time," he says — often more than a year in his experience.

Yet another challenge across all life-science fields is moving university research into biotech spin-offs. "There are still some wrong views about venture capital and scientists and making money," says Dario Grattapaglia, a forestry researcher at the Brazilian agricultural research corporation Embrapa, the science arm of Brazil's Ministry of Agriculture. Grattapaglia, who develops molecular breeding techniques to speed tree growth and select for different wood characteristics, says that Embrapa works with companies more often than does the typical university. He often collaborates with former students who now work at forest product companies.

At Fiocruz, a centre for developing health technology has been established, with a new building slated for completion in 2011 or 2012. "We're looking to establish something like incubators," says Claude Pirmez, Fiocruz's vice-president for research and reference laboratories. She anticipates that this will mean new products or perhaps research leading to clinical trials. But bureaucracies persist, and negotiating

intellectual-property transfer can be difficult even at Embrapa. Grattapaglia fondly recalls his days at North Carolina State University where he completed his PhD, acquiring two patents based on his thesis paper alone.

Nonetheless, university research can lead to promising biotech ventures. Alellyx Applied Genomics, for example, is a growing biotech born of successful genomics work in 2002. Alellyx co-founder Paulo Arruda had been a plant molecular biologist at São Paulo's University of Campinas for 30 years. His foray into biotech began in 2000 after co-authoring a paper on the first genome sequence of a plant pathogen — which was also the first

sequencing project led by Brazilian scientists (A. J. G. Simpson *et al.* *Nature* **406**, 151–157; 2000). A network of 25 labs sequenced the bacterium *Xylella fastidiosa*, which attacks citrus trees. Located in the suburb of Campinas outside São Paulo, Alellyx initially sprang up to find applications for the sequence. But, says Arruda, plans shifted in 2003 with the growing recognition of the opportunity presented by the rapidly expanding sugarcane-ethanol fuel industry. This, they calculated, was their route to profitability. Alellyx is hunting for genes to improve growth of different sugarcane varieties, and its next-door sister company, CanaVialis, breeds and markets different varieties. The two companies showed enough promise to catch the eye of US agribiotech giant Monsanto, which bought them in 2008 for \$290 million. But Alellyx is a relatively rare case. "The problem is we don't have an entrepreneurship culture in students," says Arruda, who suggests that universities should offer more business courses for science students and more internships at companies.

But neither increasing the number of research opportunities nor boosting the role of the private sector seems an insurmountable obstacle for a country that just secured South America's first-ever Olympic Games. If the industry-academia stigma recedes and salaries ascend, and native talent returns home, Brazilian life sciences will have plenty of promise — and probably plenty more temples of science.

Gene Russo is Editor of Naturejobs.

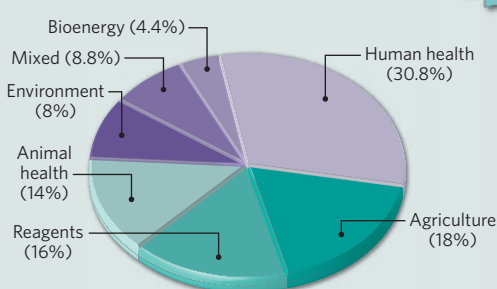


Gerardo Mendoza: bureaucracy is hampering progress.

GROWTH INDUSTRY

Brazil's burgeoning life-sciences sector.
Total number of companies: 253

LIFE-SCIENCE COMPANIES BY SECTOR



Helpdesk

Personal service.

rpg

It's the smells I find most evocative. They take me places, short-circuit the other senses. A hint of rose; elderflower in the garden. The river and its white-sanded estuary; the crashing of surf. Smoke from the soldering iron; the high-school physics lab. The softness of her belly and a warm summer's evening.

Soldering iron?

I blink my eyes, trying to focus on the swimming greenness around me. I am on a forest floor, dappled sunlight picking out — no, that's not right. On a river bed perhaps, light refracting oddly and glinting off dust motes — that doesn't work, either. Besides, I can breathe. The air, and it is definitely air, I think I would know if my lungs were full of water, is clean and sweet, but with a definite odour of electronics. Dead electronics.

My nose itches. I move my hand to — I try to move my hand to scratch it, but something isn't quite right. My arm won't move. I try to sit up, see what's happening: but my body is just as stuck. I wonder if I'm drunk. That might explain why I can't remember what happened just before I ended up ... wherever here is.

But here, here something appears. I want to say 'swims into vision', but that's not quite right, either. Nothing is quite right. There are two somethings, now. At least I think there are two. I can't focus properly. Maybe I can focus, and they are meant to be that fuzzy?

They are dark. I want to say they are the size of dinner plates, but I can't tell how far away they are. And now they shrink. Are they shrinking, or are they —

Ah. They are moving farther away. And they are framed by something that's shaped almost like a — and they're gone, and now they're back. Is that a nose, do you think? A slit in the greenness appears, a dark slit that seems to oscillate (and there again is the memory of the electronics bench swimming just out of my depth) and some strange noise in my ears.

And understanding in my head.

It is awake.

The words make themselves understood, even though I don't remember hearing them. I am surprised that I am lucid enough to realize this. But I've seen this apparition before: my heart suddenly racing, I open my mouth to scream but before I can it fills with something

— something that feels like candy floss, squeaky like a balloon; dry like the taste of chalk. I try to sit up, buttocks clenching, chest straining against the straps holding me down, pin pricks of sweat on my brow: one drops into my right eye and I can't wipe it away.

There will be a moment of readjustment.

I fall back, breathing heavily. I remember getting out of a ski lift somewhere in the Swiss Alps, fighting to draw oxygen into my lungs. But there was the hush of freshly fallen snow; here this strange, pervasive, persistent borborygmus.

It is all right. Everything will be okay.

It's not a voice, it's a certainty in your mind. Force yourself to look at the creature (naming it tames the terror even as it engenders it) and focus on it. Take in its round eyes; the holes for nostrils; the lipless, quivering mouth. Force yourself to stay still as three long, thin stalks that you suddenly know to be fingers brush lightly (oh so lightly, like the touch of a hesitant lover — but that is not the source of the smell of the solder; this is the scent of evening) over your face and remove the gag.

We crashed. There is a problem.

There is a noise, a real, honest-to-God, human noise, like the release of pressure from a train's brakes. Light (and can this be real sunlight fording this turbid air like a frontiersman?) breaks in, and the creature seems smaller somehow; no less inhuman but not, somehow, as alien.

"I'm still on Earth?" I manage to force out.

Yes. There is no structural damage. We need your help.

A certainty. But why me? I'm a programmer, not an engineer.

"Show me."

The glaucous light changes then, flickers, moves. I am on a gurney and they are transporting me ... where? I am being raised: my feet come into view, and suddenly my arms are free. Straps yet restrain my legs, but the tightness across my chest is gone. In front of me, it looks like nothing so much as a TV screen, or computer monitor. Blue screen, with white, alien characters. But some, I realize as my blood suddenly pounds in my ears, some I recognize:

0x0001000B 0x5043 ...

From the endless depths of space they came: technology to conquer distances that can only be measured in terms of photons; a civilization I can't imagine. And yet ...

I look at the blue screen again. Over my coughing, the pain in my chest, the tears that are suddenly streaming down my cheek, I hear them say, trembling, almost apologetic:

It just stopped. We didn't do anything different.

rpg: poet, scientist, gadfly. Currently rocking the world of information architecture after an enforced sojourn in the Antipodes. Nowhere is safe. Join the discussion of Futures in Nature at go.nature.com/QMAm2a



JACEY